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

index is provided for the benefit of the worker who wishes to use the book primarily for reference.

The material in this book is largely an integration of the results of several papers I have published during the last decade. I have also attempted to relate this material to similar material recently published by other authors. Many of these latter references are to books and articles appearing during the last five years. I have not attempted to study the historical background for every method in sufficient detail to enable me to state who was the first author of it, but, frequently, I have given historical information that seems to be of general interest.

Special emphasis is given in this book (see Chapters 2 and 17) to the general subject of the accumulation of errors when the computations involve approximate numbers.

The notation of my book, which differs in some respects from the notation used by other writers and, sometimes, from the notation I have used in previous work, has been selected carefully and, in some cases, after considerable consultation. The most drastic change from past practice deals with the order of the primary subscripts in the indicated solution. This order is the reverse of that of the Yule notation, now quite universally accepted by statisticians.

Several of my colleagues have assisted in reading the manuscript, or certain portions of it, and have made valuable suggestions for its improvement. In particular I wish to express my appreciation to Carl A. Bennett, Professor George M. McEwen, Jack Northam, Dr. C. V. Winder, and to many former students who have responded to the material as it was presented in class. My chief acknowledgment, however, is to my wife, who gave so much care to the preparation of the manuscript.

I hope that this book will aid the research worker not only in supplying him with definite techniques for specific problems but also, more generally, in designing appropriate calculational techniques through the use of better methods that feature less recording, more adequate checking to control mistakes, better control of errors, and relative ease of computation.

PAUL S. DWYER

University of Michigan
April, 1951

Contents

CHAPTER 1

Introductory Remarks

1.1 Introduction. Computation, before the invention of the modern calculating machine, consisted in the successive applications of the fundamental operations of addition, subtraction, multiplication, and division. The modern calculating machine, as it was gradually perfected, enabled the computer to perform these fundamental operations much more easily, quickly, and accurately. It is not surprising, then, that the utility of the machine in performing these individual operations received wide recognition, but that the greater usefulness of the machine in integrating a series of these fundamental operations into a more inclusive operational unit was not so generally recognized. It is the purpose of this book to show how a design based on these operational units can be used in simplifying the calculational steps in many complex problems by demonstrating how the basic operations can be integrated into the design.

In particular the application of these methods in this book is made to certain linear problems. These are problems that are associated with the general question of solving sets of simultaneous linear equations. The methods of solution are sufficiently complicated to demand special attention to techniques. Many of the different computational devices presented have application to other problems, which are not essentially linear, so that this book, or at least portions of it, should have value for those who are interested in computation in general.

The term, *linear computations*, as used throughout this book, applies to the computations involved in the solution of simultaneous linear equations and such allied problems as the evaluation of determinants, the calculation of the adjoint and the inverse of a matrix, and the solution of problems involving the characteristic equation. Although the term can have a more general meaning, the treatment here is limited to linear algebraic equations, and special emphasis is given to the solutions with direct pivotal condensation methods that are particularly applicable to the modern desk calculator.

1.2 Computation with digital methods. The discussion is limited to computations involving *digital numbers*. These are numbers that can be written in terms of a finite number of digits. Thus the right side of

$$\frac{4}{3} = 1.333\cdots$$

is not a digital number in the decimal system, but the approximate value, 1.33333, is.

The modern desk calculator is a *digital machine* since the numbers must be placed in digital form before computation. It is impossible to use $\sqrt{2}$ in calculation with these machines, but some approximation to $\sqrt{2}$, such as 1.414, may be used. Punched card computers and the recently developed high-speed electronic calculators that use the binary system are likewise digital. Moreover, the operations of arithmetic feature the combination of digital addition, subtraction, and multiplication tables into suitable calculational form. Computation with common logarithms is digital. The slide rule, in contrast, is essentially non-digital, although it is used conventionally to work problems expressed in digital form, and digital results are read. Here the basic result is some length on a scale, and this length, when measured in terms of the variable units of the scale, is not necessarily digital. Other non-digital devices include the weighing machine used by the grocer, upon which weight and cost are indicated by a position on a continuous scale, and electric machines, on which the quantities to be computed are also shown by the position of an indicator on a continuous scale. Non-digital computing machines are called *analogue computers* [A].*

One cannot necessarily represent a number exactly with either a digital or an analogue computer. Analogue computers demand the transference of a number to some quantity such as distance, electric current, weight, and such transference cannot in general be exact. For example, the digital number 1.7183 cannot, in practice, be located exactly on a scale.

Computations with digital numbers do not always result in digital numbers. The results of addition, subtraction, and multiplication of digital numbers are digital, but the digital answers for division and square root are, in general, only approximate. In pure digital computation, we should avoid these latter operations or at least postpone them as long as possible.

1.3 The fundamental operations with a computing machine. Detailed instructions regarding the fundamental operations are not presented here. Such instructions, appropriate to each machine, can be

* Capital letters in brackets refer to the bibliography at the end of the chapter.

obtained from the manufacturer, and can also be found in available published material [B]. The reader should familiarize himself with the basic instructions appropriate to his particular machine.

A modern computing machine is a machine that performs addition, subtraction, multiplication, and division. Many adding machines should not be thought of as computing machines, since they cannot be used efficiently to perform multiplication and division.

In general, the methods suggested in this book are applicable to desk machines of all kinds, although they are more satisfactory with fully automatic machines. *Hand machines* are machines in which the various operations are accomplished by the turning of a crank. *Semi-automatic machines* are similar to hand machines except that the operations are carried on electrically. Most semi-automatic machines are equipped with automatic division that enables us to find, as a single operation, the decimal approximation to any ratio. *Fully automatic machines* have the general features of the semi-automatic models and the additional feature of automatic multiplication. They are equipped with one or more such features as automatic clearance, automatic return to a fixed decimal point, or automatic subtraction of products from a previously recorded total.

There is considerable variation in the design of machines, so that it is improper to use such terminology as keyboard or lower dials. An acceptable and quite universally applicable terminology [C] designates the *setting mechanism*, through which the numbers are introduced into the machine; the *revolutions register*, which records the number of turns of the crank or shaft; and the *products register*, which indicates the result of a multiplication. In division the numerator is placed in the products register, the denominator is placed in the setting mechanism, and the quotient appears in the revolutions register, with the remainder in the products register.

In some fully automatic machines a different setting mechanism is used for multiplicand and multiplier. We might designate the setting mechanism used for addition and subtraction as the *primary setting mechanism* and the additional mechanism used in multiplication as the *secondary setting mechanism*.

A machine chosen for extensive computation should be capable of performing the basic operations without mistakes. A common deficiency, even in automatic models, is inadequate carry over in the products register. Another common defect is the omission of an extra digital position at the left of the products register to take care of situations in which the result of a full-capacity operation is negative. These deficiencies are not so serious in hand and semi-automatic models where

the computer should take an active part in the details of each operation, but they are serious in a fully automatic model that should give correct results if the computer manipulates the proper keys, bars, and levers. These deficiencies are especially apparent to the scientist when he carries his calculations to machine capacity.

The inability of the machine to express negative numbers conventionally is at present one of the limitations of computational work with desk computers. When a number $b > a$ is subtracted from a, the digital process built into the machine causes 9's to appear to the left in the products register. Thus $189 - 742$ appears as 447 preceded by a series of 9's. If the machine has complete carry over, that is, if the products register operates as a unit, the 9's are present to the extreme left of the register. If, however, the machine does not have complete carry over, they are present only as far as the carriage is in mesh with the mechanism in the machine proper.

Various devices are available for obtaining the usual form of a negative number. One method is to ascertain it mentally, as the complement of the number in the products register, punching it on the primary setting mechanism as ascertained. The addition of this to the number in the products register clears the register and verifies the correctness of the result, whereas a second addition of it exhibits the absolute value of the negative number in the products register.

A second method calls for placing in the setting mechanism the number (say of k digits) that appears at the right of the 9's. One subtraction replaces this number with k zeros; a second subtraction produces the absolute value of the desired negative number in the k places at the right of the products register.

1.4 The extraction of square root. The extraction of square root is not usually considered to be one of the fundamental operations, but it is frequently required in computational work. Since it is used extensively in one of the recommended methods for solving simultaneous linear equations, a brief outline of different methods of taking square root with a modern computing machine is presented.

The calculator may serve as a mechanical supplement in the extraction of square root by any of the following methods: the digital method taught in arithmetic, the method of logarithms, the method that uses the differential, Newton's method, Horner's method, Vieta's method, expansion of series, and interpolation when tables of square roots are used. However, square root is usually extracted with machines by the subtraction of odd numbers or by methods of successive approximation with the use of division.

The justification of the odd number method is based on the fact that

the sum of the first N odd numbers is N^2. This may be written in the form $\sum_{x=1}^{N} (2x - 1) = N^2$ and results from adding the N equations

$$x^2 - (x - 1)^2 = 2x - 1 \tag{1}$$

with $x = 1, 2, 3, \cdots, N$. If the number whose square root is desired is a perfect square, it is necessary only to subtract the odd numbers, beginning with 1, since the number of the odd numbers subtracted equals the square root of the perfect square. Thus $\sqrt{25} = 5$, since $25 - 1 - 3 - 5 - 7 - 9 = 0$, and the five odd numbers are subtracted. The calculation could be designed to show 5 in the revolutions register of the machine if these subtractions were performed on a machine.

However, if a number whose square root is desired is not a perfect square, this process can be used to find a digital number whose square differs from the number by less than any predetermined amount. Thus $\sqrt{15130}$ can be approximated by 123 with an error of unity in its square since $\sum_{x=1}^{123} (2x - 1) = 15129$. Consequently $\sqrt{1.5130}$ can be approximated by 1.23, with an error in its square of 0.0001.

When the square root of a number with more than two digits is desired, we may combine the subtraction of odd numbers with digiting. Since the first ten odd numbers total 100, they may all be subtracted by subtracting 100. A "1" (representing 10) is thereby recorded in the revolutions register. The eleventh odd number, 21, is subtracted and the process proceeds. Thus $\sqrt{169} = 13$, since $169 - 100 - 21 - 23 - 25 = 0$. Now by the principle enunciated it is evident that, since in subtracting 100, $400 = 100 + 300$, $900 = 100 + 300 + 500$, $1600 = 100 + 300 + 500 + 700$, etc., we are subtracting the first 10, 20, 30, 40, etc., odd numbers. The next odd numbers are respectively 21, 41, 61, 81, etc.

These facts are the basis of the generally known technique whereby the carriage is placed as far to the right as possible, N is placed at the left of the products register, the digits are arranged in groups of 2 as measured from the decimal point, the successive odd numbers are subtracted in the group at the left until negative numbers are introduced in the products register, the last odd number is added to make the result positive, the even number just below is punched before the carriage is shifted and the subtraction of the odd numbers in the next column is started. This process is continued. If the original number is placed at the left portion of the products register, the revolutions register records the successive digits of the square root. The correct decimal point is then easily added.

The calculating machine is useful in extracting square root by iterative methods, and especially so with automatic or semi-automatic machines. The general plan is to take some approximation to $\sqrt{N}$ and to divide it into N. The average of the first approximation and the quotient is the next approximation. The process continues until stability is established to the required number of digits. The general scheme is based on the formula

$$(2) \qquad \frac{N}{\sqrt{N}+\epsilon} = \sqrt{N}\left[1+\frac{\epsilon}{\sqrt{N}}\right]^{-1} = \sqrt{N}\left[1-\frac{\epsilon}{\sqrt{N}}+\frac{\epsilon^2}{N}-\cdots\right]$$

so that the average of $\sqrt{N}+\epsilon$ and $N/(\sqrt{N}+\epsilon)$ is

$$(3) \qquad \sqrt{N}\left[1+\frac{\epsilon^2}{2N}-\cdots\right].$$

This is a satisfactory formula in that the series is an alternating series if $\epsilon > 0$, with the first-order error term not present.

A method based on successive applications of (3) rather than on the terms of the series is more satisfactory. In this case take the first approximation less than $\sqrt{N}$ rather than greater than $\sqrt{N}$ since the two results can then be averaged more easily even though the series does not alternate. Thus $2/1.4 = 1.429$ and the new estimate is $1.4 + \frac{1}{2}(0.029) = 1.4145$, whereas $2/1.5 = 1.333$ and the new estimate is not so easily obtained. The process converges more rapidly if first estimates are accurate. Such estimates may be obtained by using a brief table of square roots or a slide rule. If the first estimate of $\sqrt{2}$ is 1.414, the process gives $2/1.414 = 1.4144272$ and $\sqrt{2} = 1.4142136$.

Another successive approximation scheme makes use of the formula

$$(4) \qquad \frac{(N)+(N+\epsilon)}{2\sqrt{N+\epsilon}} = \frac{2N+\epsilon}{2\sqrt{N+\epsilon}}$$

$$= \frac{N\left(1+\frac{\epsilon}{2N}\right)}{\sqrt{N}\left(1+\frac{\epsilon}{N}\right)^{1/2}}$$

$$= \sqrt{N}\left(1+\frac{\epsilon}{2N}\right)\left(1+\frac{\epsilon}{N}\right)^{-1/2}$$

$$= \sqrt{N}\left(1+\frac{\epsilon}{2N}\right)\left(1-\frac{\epsilon}{2N}+\frac{3\epsilon^2}{8N^2}-\cdots\right)$$

so that finally

$$\frac{N + (N + \epsilon)}{2\sqrt{N + \epsilon}} = \sqrt{N}\left[1 + \frac{\epsilon^2}{8N^2} - \cdots\right]. \tag{5}$$

This formula is satisfactory since the first-order error term is again lacking and the series alternates. Values of ϵ for each term can be chosen so that the error of $\sqrt{N}$ computed by the left of (5) will not be larger than any specified amount. The manufacturers of the Marchant calculating machine have provided tables of $2\sqrt{N + \epsilon}$ and $1/2\sqrt{N + \epsilon}$, which are useful in finding $\sqrt{N}$ to five digits with one division. If more digits are desired, the method of approximation indicated in (3) can be used.

The second method, which demands a table for effective use, is not necessarily to be preferred to the first. The computer should understand each method and make his own selection.

The general methods of (2) and (4) can be extended to cube root and roots of higher order [D], [E].

1.5 Additional aids to computation. Considerable space might be devoted here to other aids to linear computation. Reference is made to several of these in [F].

REFERENCES

A. For a discussion of mathematical machines covering digital machines and analogue computers, see

1. F. J. Murray, *The Theory of Mathematical Machines*, King's Crown Press, New York, 1947.
2. The Publications of Office Machines Research, Inc., also contain much interesting information about different machines. International Office, International Research Corporation, Amsterdam, Holland.
3. Engineering Research Associates, *High Speed Computing Devices*, McGraw-Hill Book Co., New York, 1950.

B. In addition to the instructions provided by the manufacturers, descriptions of American machines and detailed instructions regarding their fundamental operations are available in

1. Katherine Pease, *Machine Computation of Elementary Statistics (with Special Reference to Fridén, Marchant, and Monroe Calculating Machines)*, Chartwell House, New York, 1949.
2. F. A. Willers, *Practical Analysis (Graphical and Numerical Methods)*, translated by Robert Beyer, Dover Publications, New York, 1948.

The first of these books contains lists of problems as well as a bibliography of manuals and methods published by calculating companies. The second contains a section, pp. 45–73, on calculation with machines, written by Tracy W. Simpson. This section concludes with a brief bibliography of

material prepared, for the most part, by the Marchant Calculating Machine Company.

C. A discussion of appropriate terminology may be found in references [A.2] and [B.2, pp. 45–50] and in

1. L. J. Comrie, "On the calculation of correlation coefficients with modern calculating machines, I," *Journal of the American Statistical Association*, **36**, 429 (1941).
2. P. S. Dwyer, "On the calculation of correlation coefficients with modern computing machines, II," *Journal of the American Statistical Association*, **36**, 429–430 (1941).

D. Additional material on the second iterative method for roots, which is recommended by the Marchant Calculating Company, is

1. *Square Root Multipliers* (for Calculating Square Root to 5 Significant Figures), Marchant Calculating Machine Co., Oakland, Calif., 1940.
2. *Square Root Divisors* (for Calculating Square Root to 5 Significant Figures), Marchant Calculating Co., Oakland, Calif., 1940.
3. *Cube Root Divisors* (for Calculating Cube Root to 5 Significant Figures), Marchant Calculating Co., Oakland, Calif., 1944.

These tables of square root divisors and cube root divisors are also available in [B.2, pp. 69–72].

E. Professor E. E. Ingalls has called my attention to the fact that iterative methods were used in extraction of cube root in early American arithmetics. For example, the *New and Complete System of Arithmetic*, by Nicolas Pike (third edition revised by Nathaniel Lord), which was published by Thomas and Andrews at Boston, in 1808, described the following iterative method on page 188. Let y^3 be the number whose cube root, y, is desired. Let x be an approximation to y. Form

$$\begin{array}{r|l} x^3 & y^3 \\ 2x^3 & 2y^3 \\ 2x^3 + y^3 & 2y^3 + x^3 \end{array}$$

and use

$$\frac{(2y^3 + x^3)x}{2x^3 + y^3}$$

as the new approximation.

F. Brief treatments of additional calculational devices such as slide rules and nomograms are available in [B.2, pp. 14–45]. Tables of addition and subtraction logarithms and theory for their use are available in

1. W. M. Johnson, *Addition and Subtraction Logarithms to Seven Decimal Places*, Charles T. Powner Co., Chicago, Ill., 1943.
2. L. M. Berkeley, *Addition-Subtraction Logarithms to Five Decimal Places*, White Book and Supply Co., New York, 1930.

A discussion of the use of punched cards is available in

3. W. J. Eckert, *Punched Card Methods in Scientific Computation*, The Thomas J. Watson Astronomical Computing Bureau, Columbia University, 1940.

Considerable insight into the status of the large-scale digital computers may be obtained from

4. *Proceedings of a Symposium on Large Scale Digital Calculating Machinery,* Harvard University Press, Cambridge, Mass., 1948, as well as from the section on automatic computing machinery that appears in the issues of [F.5].
5. *Mathematical Tables and Other Aids to Computation,* National Research Council, Washington, D. C.

EXERCISES

1. Perform the indicated operations with a computing machine and record the exact answers.

(*a*) $1.234 + 2.375 - 0.327 - 1.111$
(*b*) $(8.123)(3.456) = 8.123 \cdot 3.456 = 8.123 \times 3.456$
(*c*) $(0.00234)(6.738)$
(*d*) $(2.236 \cdot 10^{-10})(3.468 \cdot 10^{-20})$

2. Perform the indicated operations with a computing machine and record the answers to six decimal places.

(*a*) $(0.345)(1.367529)$
(*b*) $0.345 \div 1.367529$
(*c*) $1.367529 \div 0.345$
(*d*) $4 \div 7$

3. Use a calculating machine and obtain the results in suitable form.

(*a*) $169 - 432$
(*b*) $2.033 + 4.027 + 8.009 - 10.002 - 8.887$
(*c*) $(432)(-163)$

4. Calculate to four decimal places, using the method of subtraction of odd numbers. Check the answers by squaring them.

(*a*) $\sqrt{2}$
(*b*) $\sqrt{0.00329}$
(*c*) $\sqrt{16.029}$
(*d*) $\sqrt{0.09327}$

5. Obtain the answers to the various parts of exercise 4 to seven decimal places by using the results of exercise 4 as the basis of the iterative process. Check the answers by squaring them.

6. Obtain approximations to the square root of each of the following numbers by dividing each by some estimate of its square root. Keep up the process until you are certain that the result is correct to three decimal places. Check the answers by a final multiplication.

(*a*) 0.679
(*b*) 0.0679
(*c*) 7.099
(*d*) 70.99

7. Use (1.4.5) in obtaining $\sqrt{501}$ if the value of $2\sqrt{500}$ is available and is tabulated as 44.72135955. Calculate the value of $\sqrt{501}$ to eight decimal places by one of

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 $3(N + \epsilon)^{\frac{2}{3}}$.

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

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

and

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

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 $6.738(\frac{1}{2})$ 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

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

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

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

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

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

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

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

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

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

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

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

$$\begin{bmatrix}1.235\\1.225\end{bmatrix}\begin{bmatrix}2.345\\2.335\end{bmatrix} = \begin{bmatrix}2.896075\\2.860875\end{bmatrix}.$$

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

$$\begin{aligned} \frac{x_1}{x_2} &= \frac{x'_1 + \epsilon_1}{x'_2 + \epsilon_2} \\ &= \frac{x'_1\left(1 + \dfrac{\epsilon_1}{x'_1}\right)}{x'_2\left(1 + \dfrac{\epsilon_2}{x'_2}\right)} \\ &= \frac{x'_1}{x'_2}\left(1 + \frac{\epsilon_1}{x'_1}\right)\left(1 + \frac{\epsilon_2}{x'_2}\right)^{-1} \\ &= \frac{x'_1}{x'_2}\left(1 + \frac{\epsilon}{x'_1}\right)\left(1 - \frac{\epsilon_2}{x'_2} + \frac{\epsilon_2{}^2}{x'_2{}^2} - \cdots\right) \\ &= \frac{x'_1}{x'_2}\left(1 + \frac{\epsilon_1}{x'_1} - \frac{\epsilon_2}{x'_2} + \cdots\right). \end{aligned} \tag{7}$$

* 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

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

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

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

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

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

whereas, if x_2 is an exact number, it becomes

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

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 $x' > 0$, $d(\sqrt{x'}) = \dfrac{1}{2\sqrt{x'}}\,dx'$, 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

$$\begin{aligned} \log_e P &= \log_e x + \log_e y \\ \log_e Q &= \log_e x - \log_e y \\ \log_e U &= p \log_e x \\ \log_e V &= \frac{1}{p} \log_e x. \end{aligned} \tag{2}$$

It follows that

$$\begin{aligned} \frac{dP}{P} &= \frac{dx}{x} + \frac{dy}{y} \\ \frac{dQ}{Q} &= \frac{dx}{x} - \frac{dy}{y} \\ \frac{dU}{U} &= p\frac{dx}{x} \\ \frac{dV}{V} &= \frac{1}{p}\frac{dx}{x}, \end{aligned} \tag{3}$$

which gives us

$$|\epsilon_r(P)| \leq |\epsilon_r(x)| + |\epsilon_r(y)|$$

$$|\epsilon_r(Q)| \leq |\epsilon_r(x)| + |\epsilon_r(y)|$$

$$|\epsilon_r(U)| \leq p\,|\epsilon_r(x)| \tag{4}$$

$$|\epsilon_r(V)| \leq \frac{1}{p}|\epsilon_r(x)|.$$

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

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

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

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

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

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

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

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

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* $\epsilon_r(x) \leq \dfrac{1}{(k+1)10^{n-1}}$, *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* $x \leq \dfrac{1}{2(k+1)10^{n-1}}$, *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 ease 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)

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

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

$$(1) \quad \epsilon_r(x_1 x_2) = \epsilon_r(I_1 I_2) = \frac{1}{2k_1 \cdot 10^{n-1}} + \frac{1}{2k_2 \cdot 10^{n-1}}$$

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

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 $3 = \left|\begin{matrix}3.5\\2.5\end{matrix}\right|$. The use of approximation-error numbers in 2.8 shows that a much better answer is $2.878(18) = \left[\begin{matrix}2.896\\2.860\end{matrix}\right]$.

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*) $63.9 \cdot 10^{-8}$
(*c*) 92,500,000
(*d*) $62.5 \cdot 10^{20}$

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.

CHAPTER 3

The Principles of Computational Design

3.1 Introduction. A calculational method should be designed to make maximum use of the available knowledge and equipment. Such a method is herein termed "good," and the principles that lead to its construction are called "the principles of computational design."

Whether a method of calculation may be termed "good" depends, of course, on the purpose of the calculation. A method that is good for obtaining approximate answers quickly may not be a "good" method for getting exact answers. A sound computational method can be devised only if we begin with a clear realization of the purpose of the calculation and then select the combination of those principles that can best attain the end. Not every one of the principles enunciated in this chapter should be used in solving any specific problem; however, the whole computational procedure should be directed toward the desired end with full consideration of all pertinent knowledge, techniques, and flexibility of equipment. The remaining sections of this chapter are devoted to the problem of computational design with the use of mechanical equipment and, specifically, with a modern desk computer. Some suggestions may have use, however, in connection with other computing devices.

3.2 Criteria for a good computational method. The following are suggested as criteria for a good computational method:

(*a*) The method should provide for control of mistakes. If mistakes are made, the method should make possible their discovery and elimination.
(*b*) The method should make possible the control of errors.
(*c*) The method should provide for adequate checking and, when possible, for final verification.
(*d*) The method should feature a minimum of recording and resetting.
(*e*) The method should demand a minimum of time for its execution.
(*f*) The method should be carried out with relative ease.

The first three criteria may be considered to be of a somewhat theoretical nature, since they deal with the accuracy and precision of the result. The last three criteria are more practical in nature, being concerned in decreasing the amount of time and energy spent and the difficulty of the computation. Better computational design, as judged by the last three criteria, is usually better computational design as judged by the first three, since the ease of calculation and a minimum of recording tend to cut down the mistakes, while the smaller amount of time required makes possible additional checking and verification.

In so far as error control is concerned, we should note four special types of problems. They are:

(a) The exact solution of exact problems.
(b) The approximate solution of exact problems.
(c) The exact solution of approximate problems.
(d) The approximate solution of approximate problems.

In type a there is complete error control since no error may enter the computation. In type b the solution may be obtained theoretically to any degree of accuracy, that is, the errors may be made smaller than any specified quantity. Type c at first seems a bit academic, but, if we treat the approximate numbers as though they were exact and if we use exact methods, the results are exact from a calculational standpoint. In this way all error terms are eliminated and we can concentrate on the elimination of mistakes. We should be careful, however, not to interpret the final result of the calculation as having no error.

As an illustration of the exact solution of approximate problems, let us consider two linear equations in two unknowns with the coefficients subject to error. For example, we desire the values of x_1 and x_2, which satisfy

$$\begin{aligned}(a_{11} + \epsilon_{11})x_1 + (a_{12} + \epsilon_{12})x_2 &= a_{13} + \epsilon_{13} \\ (a_{21} + \epsilon_{21})x_1 + (a_{22} + \epsilon_{22})x_2 &= a_{23} + \epsilon_{23}.\end{aligned} \tag{1}$$

If the ϵ's, as well as the a's, are known, it is possible, in general, to get a unique exact solution for x_1 and x_2. But usually the ϵ's are not known. We know simply that the coefficients are approximate numbers, each having a specified range. We may then write (1) in the form

$$\begin{aligned}a_{11}x_1 + a_{12}x_2 &= a_{13} \\ a_{21}x_1 + a_{22}x_2 &= a_{23}\end{aligned} \tag{2}$$

with the understanding that each a_{ij} is an approximate number that may take on an infinity of values. From this standpoint (2) represents

not one pair of equations, but a sextuple infinity of such pairs, each of which presumably has a unique solution. We may define the solution of all these as the exact solution of (2), with the a_{ij} treated as exact numbers, if we understand that this exact solution is the exact solution of (2) when the a_{ij} are exact, but that it is an approximation to any of the infinity of solutions obtained if the a_{ij} are approximate. It is shown later that bounds for the errors of the solutions may be obtained if bounds for the errors of the coefficients are known.

A final verification consists in showing that the proposed answer satisfies all the conditions of the problem. In a trivial illustration we verify that the solution of $3x = 12$ is 4, since three times four equals twelve. Final verification of this sort is not always feasible, but it can be used quite generally in connection with the solutions of simultaneous linear equations.

3.3 Use of checking devices. Checking devices should be introduced regularly in connection with extensive numerical work. These may be of a direct nature, such as repeating an operation, or, preferably, of an indirect nature, such as using a row sum check. Sometimes an alternative computational procedure serves as a check.

In some cases mathematical devices can be used. The row sum, or column sum, check, which is designed to utilize a mathematical property, is a device of this sort. The method of multiplication and subtraction with exact division (Chapter 5), based on the mathematical fact that each recorded result is obtained from an exact division of the preceding pivot, is an illustration of the use of a mathematical property.

3.4 Avoidance or postponement of approximate operations. In case exact methods are being used, approximate operations (division or square root) should not be used; or, if it is impossible to eliminate them entirely, they should be delayed as long as possible. Even in calculations with approximate numbers, for purposes of checking, the exact operations of addition, subtraction, and multiplication should be used in preference to approximate operations.

It is the author's thesis that *approximate operations should be deferred as long as possible.* Two illustrations are presented to support this.

The conventional procedure used in calculating the sum of the squares of the deviations in connection with the technique of the analysis of variance is the use of the formula

$$\Sigma x^2 = \Sigma X^2 - \frac{(\Sigma X)(\Sigma X)}{N} = \Sigma X^2 - (\Sigma X)(\overline{X}), \tag{1}$$

where $\overline{X} = \Sigma X/N$. After the values N, ΣX, and ΣX^2 are obtained,

this technique calls for the (usually) approximate operation of division as the first step. This is in direct violation of the principle stated above, which indicates that the division should be deferred. Now this division can be deferred until the last operation of the calculation with the use of the calculational formula,

$$\Sigma x^2 = \frac{N\Sigma X^2 - (\Sigma X)^2}{N}. \tag{2}$$

Formula (2) is theoretically superior to (1) if N, ΣX, and ΣX^2 are exact in that the result can be computed with a single machine operation, whereas, in general, it is impossible to obtain the result by (1) exactly after an approximation has been made for $\bar{X}$. Thus if $N = 19$, $\Sigma X = 1272$, $\Sigma X^2 = 89704$,* formula (2) yields $\Sigma x^2 = 4546\frac{18}{19}$, whereas the use of (1) demands some approximation to $\bar{X}$ such as 66.95, with the resulting approximation to Σx^2 such as 4543.60. The introduction of this approximation is very unfortunate in analysis of variance work, especially since exact checks (useful in discovering and eliminating mistakes) are available. For this reason most authors recommend the recording of $\bar{X}$ to many more digits than are needed for the recording of Σx^2. This is all unnecessary as (2) gives an exact answer and enables us to round off to any desired number of places.

The formula (2) is also preferable from a practical standpoint since it can be carried out as a single machine operation without recording the computational entry $\bar{X}$. If X is an integer expressed to d decimal figures and ΣX^2 to $2d$ decimal figures, the value of Σx^2 is easily obtained to $2d$ decimal figures with an exact remainder. The numerator is set up on the right side of the machine as an operation involving $2d$ decimal places by multiplying the $2d$ number ΣX^2 by the integer N and subtracting the $2d$ number $(\Sigma X)^2$. The setting mechanisms and revolutions register are then cleared without clearing the products register, the integer N is set at the right of the setting mechanism, and the carriage is moved to the right. The division is then performed, and the result to $2d$ places appears in the revolutions register with the remainder in the products register. Thus in the problem above, if each unit is ten times the unit used previously,

$$N = 19, \quad \Sigma X = 127.2, \quad \Sigma X^2 = 897.04, \quad \text{and} \quad \Sigma x^2 = 45.46 + \frac{18}{1900}.$$

* These values were obtained from the series of nineteen measurements: 56, 63, 72, 68, 42, 93, 57, 44, 67, 78, 63, 42, 97, 82, 88, 71, 69, 58, 62. The nature of the unit of measurement is not important to this discussion.

The postponement of this division by N is both theoretically and practically better. In some cases it is possible to eliminate a division entirely. For example, it is possible to work out a general analysis of the large variance, $L_{xx} = N\Sigma X^2 - (\Sigma X)^2$, which parallels the usual analysis of variance based on Σx^2 but is theoretically and practically preferable in that the continued divisions by N are eliminated. See section 18.5.

The actual elimination of certain divisions is also illustrated in connection with the calculation of the correlation coefficient. A calculational formula that has been used extensively is

$$(3) \qquad \rho = \frac{\dfrac{\Sigma X_i Y_i}{N} - \left(\dfrac{\Sigma X_i}{N}\right)\left(\dfrac{\Sigma Y_i}{N}\right)}{\sqrt{\left[\dfrac{\Sigma X_i^2}{N} - \left(\dfrac{\Sigma X_i}{N}\right)^2\right]\left[\dfrac{\Sigma Y_i^2}{N} - \left(\dfrac{\Sigma Y_i}{N}\right)^2\right]}}.$$

It is now generally recognized that a better calculational formula results from eliminating the many divisions by multiplying numerator and denominator by N with the resulting

$$(4) \qquad \rho = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}}.$$

It is to be noted that the divisions by N have been completely eliminated. With the use of this formula exact answers may be had for each operation down to the square root and the final division. The final result is not difficult to obtain to any specified number of digits within machine capacity. The computational technique is better from a practical standpoint also, since it is necessary only to record the large covariance and variances, $L_{xy} = N\Sigma XY - (\Sigma X)(\Sigma Y)$, $L_{xx} = N\Sigma X^2 - (\Sigma X)^2$, $L_{yy} = N\Sigma Y^2 - (\Sigma Y)^2$ to indicate the calculational steps.

3.5 Use of indirect methods. Indirect methods may be distinctly preferable to direct methods for computational purposes. For example, the value of the third standard moment of a finite parent is given by

$$(1) \qquad \alpha_3 = \frac{\Sigma t_i^3}{N} \qquad \text{with} \qquad t_i = \left(\frac{X_i - \bar{X}}{\sigma_x}\right).$$

The formula appears simple enough at first, but the successive calculations of $\bar{X}$, σ_x, t_i, t_i^3, and $\dfrac{\Sigma t_i^3}{N}$ do not constitute an elegant computational procedure even though it is direct. An indirect method (that does not exhibit any of the deviates or standard deviates), obtained by the

mathematical substitution of the values t_i, σ_x, and $\bar{X}$ and making use of the values of section 3.4, gives us the preferable formula

$$\alpha_3 = \frac{N^2\Sigma X^3 - 3N(\Sigma X^2)(\Sigma X) + 2(\Sigma X)^3}{[N\Sigma X^2 - (\Sigma X)^2]^{3/2}}. \tag{2}$$

This formula may not at first appear to the reader to be preferable, for computational purposes, to (1), but actual computation with these formulas shows that (2) is superior, both on practical and theoretical grounds, when computing machines capable of cumulating N, ΣX, ΣX^2, and ΣX^3 are available. This situation illustrates the important fact that it is not necessarily the formula that appears most simple that is the best for computation. Formula (1) appears simple, but this simplicity is apparent rather than real since numerous auxiliary (and some of them approximate) operations are demanded before the formula can be used. Formula (2), on the other hand, appears fairly complicated, and yet it is a simpler formula for computational purposes since it enables us to obtain the result with fewer, and theoretically better, operations. An improvement of (2) is explained in the section immediately following.

3.6 Use of mathematics. Mathematics is frequently useful in calculational design. The various steps may often be carried out theoretically with the use of mathematics rather than actual computation. For example, a number of the operations implied in (3.5.1) have been carried out by mathematics to get (3.5.2). The use of additional algebra enables us to develop (3.5.2) into a computational formula that is better adapted to use with a computing machine. This formula is

$$\alpha_3 = \frac{N[N\Sigma X^3 - (\Sigma X^2)(\Sigma X)] - 2\Sigma X[N\Sigma X^2 - (\Sigma X)^2]}{[N\Sigma X^2 - (\Sigma X)^2]^{3/2}}. \tag{1}$$

Formula (1) does not at first appear to be as simple as (3.5.2), but it is much better from a computational standpoint since $L_{xx} = N\Sigma X^2 - (\Sigma X)^2$, $L_{x^2,x} = N\Sigma X^3 - (\Sigma X^2)(\Sigma X)$ and the numerator, $NL_{x^2,x} - (2\Sigma X)L_{xx}$, can be computed with operations of the type $ab - cd$.

Mathematical devices are used frequently in reducing the size of numbers by subtracting constants, dividing by constants, etc. The methods of computation of moment statistics with class interval units, which are especially effective when we do not have access to machines, serve as excellent illustrations.

Mathematics is also very useful in developing the theory that enables a simple machine to produce the results of more complicated machines. For example, a pure adding machine (such as the punched card tab-

ulator) may turn out the results of sums of products with the use of the theory of cumulative totals [A]. This device is used very extensively in computing correlation coefficients with the punched card tabulators.

In general, the computational procedure should be in mathematical form (presumably algebraic or matrix) so that the computation may be as simple as possible. Thus any expression to be calculated such as the right side of (3.4.3) should be examined critically for simplification before numerical substitution. Matrix theory is used extensively in Chapters 13 to 15 in establishing improved calculational techniques.

3.7 Use of synthetic methods. It is not necessary to record the formula to be used if it is to be used over and over again, nor is it necessary to label every term. A calculation can be put in a given position while the appearance of this calculation in this position implies that certain operations have been carried out. Thus the calculation of (3.6.1) might be placed in the form:

$$\begin{array}{llll} N & \Sigma X & \Sigma X^2 & \Sigma X^3 \\ N & 2\Sigma X & L_{xx} & L_{x^2,x} \\ & & \sqrt{L_{xx}} & NL_{x^2,x} - 2(\Sigma X)L_{xx} \\ & & L_{xx}\sqrt{L_{xx}} & \alpha_3 \end{array}$$

The calculation for the α_3 of the measures of the illustration of section 3.4, would then appear, since $\Sigma X^3 = 6{,}626{,}988$ as:

19	1272	89704	6,626,988
19	2544	86392	11,809,284
		293.92516	4,595,148
		25,392,782	0.1809628

Ordinarily we do not compute α_3 to so many places, but the method makes it possible.

Synthetic methods are used quite generally in algebra in connection with continued division. They should be used more generally in connection with other problems involving a series of operations.

3.8 Use of available mechanical features. In designing calculational techniques, the mechanical features provided by the machine should be utilized to the fullest extent. The ability of a machine to cumulate results is not always recognized. Recording individual products bc, de, and fg in the operation $a + bc + de + fg$ is inefficient since the machine can be made to record the accumulating total. Sound computational design demands the elimination of the recording of these purely incidental results unless they are needed for some other purpose.

The ability to cumulate the results is a most useful property that is usually available on even primitive mechanical equipment. The abacus, for example, as well as the modern adding machine and the modern computing machine, can cumulate results.

Another feature that should be considered in computational design with modern machines of large capacity is the possibility of simultaneous computation. It is possible, for example, to carry out different multiplications simultaneously, the design of the calculation being such that the different results appear in different portions of the products register. When this feature is combined with the cumulative feature of the machine, we are able to obtain ΣX, ΣY, ΣX^2, ΣXY, and ΣY^2 simultaneously. This design is most efficient with fully automatic machines that can square a number with a single setting. It seems unfortunate that the manufacturers of these machines have not seen the advisability of also providing a counter for N, the number of pairs.

Most machines are designed in such a way that it is easy to work with decimals, especially with a fixed number of places. It usually makes for good computational design, at least in so far as the practical aspects are concerned, to carry a fixed number of places through the calculation. Thus

$$\frac{(8.1732)(2.7318) - (4.1177)(3.2222)}{8.7769} \tag{1}$$

can be set up as a single machine operation with the result 1.0322.

If we work with a fixed number of decimal places, we must understand that we are working with incomplete numbers and not with significant numbers, for there is no guarantee that the result of a series of operations (with numbers significant to a specified number of places) is significant to the same number of places.

A fixed number of decimal places is also desirable in connection with operations involving multiplication and division with approximation-error numbers. In this case we select the fixed number of decimal places as the number of decimal places of the approximation-error number having the largest number of decimal places. Thus the value of $2.390(6) \times 4.7834(2) - 1.34(2) \times 3.333(3)$ should be written in the form

$$2.3900(60) \times 4.7834(2) - 1.3400(200) \times 3.3330(30)$$

for easy computation. The approximate value is of course

$$2.3900 \times 4.7834 - 1.3400 \times 3.3330 = 6.9661,$$

whereas the error term is

$$0.0001[2.3900 \times 2 + 4.7834 \times 60 + 1.3400 \times 30 + 3.3330 \times 200] = 0.0999.$$

The result, therefore, is

$$6.9661(999).$$

This rule indicates that operations with significant numbers having the same number of places should be carried to one more place since a possible error of one-half in the last position should be written as a 5 in the next position. Thus the calculation of the error number expressing the result of (1), if each number recorded is a significant number, appears as

$$\frac{8.17320(5) \times 2.73180(5) - 4.11770(5) \times 3.2222(5)}{8.77690(5)} = 1.03220(11).$$

Although the calculation of the approximation can be performed in a single step, the calculation of the error takes two steps. We first calculate the numerator to get

$$\frac{9.05949(91)}{8.77690(5)}$$

so that

$$R = 1.03220(11).$$

3.9 Operational units. In the past, machines have been used too frequently for the purpose of carrying out the individual operations of addition, subtraction, multiplication, division, and square root more quickly and easily. Sound computational design indicates that we should use more extensive operations, which can be accomplished easily on the machine, as the basic units of computational procedure. For example, it is shown in section 3.7 that the numerator of the formula for α_3 may be computed by a series of $ab - cd$ operations. There are other operations, listed below, that can be used similarly, in preference to the more elementary fundamental operations, as the basic units of computational design. The author defines an operation as an *operational unit* if it can be computed as a single continuous operation on the products or revolutions register. It is not necessary to watch or necessary to record the intermediate results, but the final results only. Thus in computing $ab - cd$ the values of ab and cd need not be recorded.

The operational units depend upon the equipment available. For the most part, however, modern calculating machines are flexible enough to permit a computational technique designed around the fol-

lowing operational units. These operational units are composed of combinations of the basic operations of addition, subtraction, multiplication, division, and square root.

$$U_1 = a \pm bc$$

$$U_2 = ab \pm cd$$

$$U_3 = a \pm bc \pm de \pm fg \pm, \text{ etc.}$$

$$U_4 = ab \pm cd \pm ef \pm gh \pm ii \pm, \text{ etc.}$$

$$U_5 = \frac{a}{b} \pm \frac{c}{d} \pm \frac{e}{f} \pm, \text{ etc.}$$

$$U_6 = \frac{bc}{e}$$

$$U_7 = a - \frac{bc}{e} = \frac{ae - bc}{e}$$

$$U_8 = \frac{ab - cd}{e}$$

$$U_9 = \frac{a \pm bc \pm de \pm fg \pm, \text{ etc.}}{h}$$

$$U_{10} = \frac{ab \pm cd \pm ef \pm, \text{ etc.}}{h}$$

$$U_{11} = \sqrt{a \pm bc}$$

$$U_{12} = \sqrt{ab \pm cd}$$

$$U_{13} = \sqrt{a \pm bc \pm de \pm fg \pm, \text{ etc.}}$$

The results of the first four of these operations are recorded in the products register; the rest appear in the revolutions register. In general it is possible, for each of these operational units, to enter each number to a specified number of places and to obtain the answer to this number of places.

The first four of these operational units need little discussion. However, a negative answer might appear in complement form and should be transformed to the usual negative form.

The U_5 operation can be carried out by working around a fixed decimal point, by allowing the quotients to accumulate in the revolutions

register, and by using the lever that controls the direction of the revolutions register.

The next five operations (U_6 to U_{10}) are similar division operations. It is necessary only to compute the numerator, clear the setting mechanism and the revolutions register, enter the denominator at the right of the keyboard, and move the carriage to the right sufficiently to make the division. The results are exact since the remainder is available. For example, U_9 is illustrated by

$$\frac{0.734 + (0.144)(0.733) - (0.541)(0.046) + (1.713)(0.111)}{0.423}$$

$$= 2.375 + \frac{184}{423} \text{ of } 0.001.$$

We proceed as before if the numerator is negative, except that the denominator is multiplied by 1 after the carriage has been moved to the right before the division has been started, so as to remove the 9's. The revolutions register is then made to subtract, by devices that vary with the different computing machines, and the 1000 $\cdots$ is reduced to the proper quotient. Thus

$$\frac{0.123 - (0.436)(0.327) - (0.182)(0.946)}{0.444}$$

$$= -0.432 + \frac{64}{444} \text{ of } 0.001 = -0.432.$$

Note that the quotient is negative and the remainder positive. For example,

$$\frac{(4)(2) - (10)(3)}{3} = \frac{\cdots 99978}{3} = -100 + \frac{278}{3} = -\left(100 - \frac{278}{3}\right)$$

$$= -\left(8 - \frac{2}{3}\right) = -8 + \frac{2}{3} = -7\tfrac{1}{3}.$$

In the rounding-off process, a decrease in the absolute value of the answer is called for if the remainder is more than one-half the divisor.

Perhaps the last three operations (U_{11} to U_{13}) should not be considered as operational units at all, but they can be made so with the use of successive division. If a first division by an estimated square root does not yield sufficient accuracy, it is possible to multiply the divisor by the quotient without clearing the products register to get the orig-

inal number. This is then divided by the next approximation. Tables of square root or a slide rule may be used to assist. Thus to calculate

$$\sqrt{0.904 - (-0.343)(-0.927) - (0.603)^2}$$

we first get 0.222430 in the products register. A table of square roots gives us 0.47 as a two-decimal place approximation. The division results in 0.4732+ and the three-place square root is 0.472.

Many of these operational units are used later in connection with the calculational design of the solution of linear equations.

In connection with any series of computations we should attempt to see if we can design our computation to make use of operational units [B].

3.10 Recording units. We may develop a very compact computational scheme, but the plan may involve the transfer of numbers from the revolutions register or the products register to the setting mechanism. Operations of this sort do not satisfy the definition of operational units. Since it is unnecessary to record any of the numbers in these transfers, but only the result of the series of calculations, the operation covering this procedure might be called a *recording unit*. Thus the operation *abcd* is not an operational unit with most available machines, since it is necessary to transfer results by hand from the products register to the setting mechanism. (One of the latest fully automatic machines does provide, with certain limitations, a means for doing this mechanically, so that the operation with this machine becomes an operational unit rather than a recording unit.) Sequences of operations involving additions, subtractions, multiplications, and square roots may be worked into convenient recording units.

The number of elementary operations involved in a suitable recording unit depends on the equipment available and on the abilities of the computer. The unit should be inclusive enough to cut the recording to a minimum without interfering with its essential unity. Each recording unit should be easily checked, and the checking may be difficult if this unity is destroyed.

3.11 Errors of operational units. Operational and recording units are especially to be recommended in connection with calculations using incomplete numbers. There is no guarantee that the error term is an operational unit even though the computation of the approximation term is an operational unit. The illustration of section 3.9 shows that the calculation of the error of the operational unit U_9 is not an operational unit. The calculation of these error terms may be carried out on the bases of recording units.

A useful operational unit in which the calculation of the error term is also an operational unit is

$$U_2 = ab - cd = x_1x_2 - x_3x_4$$

since the formula for $|\epsilon(U_2)|$ is

$$|\epsilon(U_2)| \leq |x_1|\eta_2 + |x_2|\eta_1 + |x_3|\eta_4 + |x_4|\eta_3 \tag{1}$$

where η_i is the maximum absolute error of x_i. For example, the error term of $(2.0)(3.0) - (1.0)(4.0)$, where the values are significant numbers, is

$$(2.00)(0.05) + (3.00)(0.05) + (1.00)(0.05) + (4.00)(0.05) = 0.5$$

so that the value is $2.0(5) = \begin{bmatrix} 2.5 \\ 1.5 \end{bmatrix}$.

Another operation whose error can be computed as an operational unit is the operation $U_8 = (ab - cd)/e$, when e is an integer. Application of (2.8.11) shows that the error of the $ab - cd$ operation can be divided by e as a single operational unit. Thus

$$\frac{(2.0)(3.0) - (1.0)(4.0)}{5} = 0.4(1) = \begin{bmatrix} 0.5 \\ 0.3 \end{bmatrix}$$

if the terms in the numerator are significant numbers and the denominator is an exact number.

REFERENCES

A. A description of the use of punched card machines in obtaining the sums of products without the individual products may be found in

1. P. S. Dwyer, "The computation of moments with the use of cumulative totals," *Annals of Mathematical Statistics*, **9**, 288–304 (1938). A fairly extensive bibliography is available here.
2. W. J. Eckert, *Punched Card Methods in Scientific Computation*, The Thomas J. Watson Astronomical Computing Bureau, Columbia University, 1940.

B. Various operational units have been described by previous writers, although the term has probably not been used previously. Some of the operational units described above are presented in

1. F. A. Willers, *Practical Analysis (Graphical and Numeric Methods)*, translated by Robert T. Beyer, Dover Publications, New York, 1948, pp. 52–68.
2. Katherine Pease, *Machine Computation of Elementary Statistics*, Chartwell House, Inc., New York, 1949.

EXERCISES

1. Calculate ΣX, ΣX^2, $L_{xx} = N\Sigma X^2 - (\Sigma X)^2$, and $\Sigma x^2 = L_{xx}/N$ if the values of X_i are 68, 70, 67, 72, 71.

2. Calculate the values of ΣX, ΣY, ΣX^2, ΣXY, ΣY^2 simultaneously and thence L_{xx}, L_{xy}, and L_{yy} if the values of (X,Y) are (8,9), (3,4), (3,5), (4,8), (4,3), and (5,5).

3. Calculate $\dfrac{(9.3284)(3.4833) - (4.2397)^2}{2.2345}$ to four decimal places as a single machine operation.

4. Find the value of the approximation-error number

$$\frac{9.32(5) \times 4.192(2) - 3.1932(3) \times 5.2(1)}{4.444(4)}$$

to four decimal places.

5. Calculate

(*a*) $0.777 - (0.333)^2$

(*b*) $0.777 + (0.333)(0.444)$

(*c*) $0.777 - (0.111)(0.222) + (0.121)(0.211) - (0.010)^2$

(*d*) $(0.888)(0.999) - (0.777)(0.666) + (0.555)(0.444)$

(*e*) $\dfrac{2.000}{1.000} + \dfrac{1.000}{2.000}$

(*f*) $\dfrac{2.000}{1.000} - \dfrac{1.000}{2.000}$

(*g*) $\dfrac{(0.743)(0.692)}{0.341}$

(*h*) $0.593 - \dfrac{(0.666)(0.742)}{0.987}$

(*i*) $24238 - \dfrac{(348)^2}{5} = \dfrac{5(24238) - (348)^2}{5}$

(*j*) $\dfrac{(0.573) + (0.123)(0.323) - (0.323)(0.427)}{0.731}$

(*k*) $\dfrac{(0.573)(0.783) - (0.123)(0.328) + (0.444)^2}{1.732}$

(*l*) $\sqrt{0.432 - (0.348)(0.412)}$

(*m*) $\sqrt{(0.432)(0.769) - (0.761)(0.011)}$

(*n*) $\sqrt{1.000 - (0.313)^2 - (0.412)^2}$

(*o*) $\sqrt{0.876 + (0.333)(0.487) - (0.421)(0.327)}$

(*p*) $\dfrac{0.369 - (0.888)(0.723) + (0.111)(0.732)}{0.678}$

to three decimal places.

CHAPTER 4

The Solution of Simultaneous Equations with the Method of Multiplication and Subtraction

4.1 Introduction. In this and the following chapters we consider methods of solving simultaneous linear equations,

$$\sum_{i=1}^{p} a_{ij}x_i = a_{i,p+1}, \qquad i = 1, 2, \cdots, p. \tag{1}$$

The general methods outlined in this and the two chapters immediately following are elimination or condensation methods in which the p equations in p unknowns are reduced to $p - 1$ equations in $p - 1$ unknowns, etc. For purposes of illustration the presentation is given in considerable detail where $p = 4$. In this case (1) becomes

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + a_{14}x_4 &= a_{15} \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + a_{24}x_4 &= a_{25} \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + a_{34}x_4 &= a_{35} \\ a_{41}x_1 + a_{42}x_2 + a_{43}x_3 + a_{44}x_4 &= a_{45}. \end{aligned} \tag{2}$$

The emphasis of the present chapter, and of those immediately following, is on the technique of solution. For this reason it is assumed that the coefficients of (1) and (2) are real and exact. Under these conditions we need only concern ourselves with the exact and approximate solutions of (1). The treatment of the situation in which the coefficients are themselves subject to error is reserved for Chapter 17 since additional material, presented in intervening chapters, is necessary for an adequate discussion of this case.

The methods of Chapter 4 and Chapter 5 are exact methods since the elimination or condensation process results in new equations with exact numbers for coefficients. The methods of Chapter 6 are approximate

(division) methods since the condensation process results in equations whose coefficients are approximate numbers. Most of the methods of Chapters 4 to 6 are pivotal methods. A *pivotal method* is one in which a particular coefficient plays a dominant role at each stage of the elimination or condensation process.

4.2 The forward solution of the method of multiplication and subtraction. The method of this chapter is called the method of multiplication and subtraction since it features the multiplication of two different rows by such multipliers as are necessary to obtain identical coefficients for some variable, followed by the subtraction of one of these equations from the other so as to eliminate this variable [A]. The common technique is to use the first equation as a means of eliminating the first variable, at each successive step, although a more general technique may be utilized. Thus we may eliminate the x_1 between the first two equations of (2) by multiplying the second equation by a_{11}, the first by a_{21}, and subtracting, with the result

$$(1)\quad (a_{11}a_{22} - a_{21}a_{12})x_2 + (a_{11}a_{23} - a_{21}a_{13})x_3 + (a_{11}a_{24} - a_{21}a_{14})x_4 = a_{11}a_{25} - a_{21}a_{15}.$$

In a similar manner we eliminate x_1 between the first and third equations of (4.1.2), and then again between the first and fourth equations of (4.1.2), and arrive at three equations in the unknowns x_2, x_3, and x_4. In a numerical problem the coefficients of the resulting equations are exact numbers. The algebraic presentation is made more compact with the use of the substitution

$$(2)\qquad m_{ij\cdot 1} = a_{11}a_{ij} - a_{i1}a_{1j}$$

so that the three equations resulting from the elimination of x_1 in (4.1.2) appear as

$$(3)\qquad \begin{aligned} m_{22\cdot 1}x_2 + m_{23\cdot 1}x_3 + m_{24\cdot 1}x_4 &= m_{25\cdot 1} \\ m_{32\cdot 1}x_2 + m_{33\cdot 1}x_3 + m_{34\cdot 1}x_4 &= m_{35\cdot 1} \\ m_{42\cdot 1}x_2 + m_{43\cdot 1}x_3 + m_{44\cdot 1}x_4 &= m_{45\cdot 1}. \end{aligned}$$

This process is a pivotal process with a_{11} the pivot, since a_{11} is used in the calculation of each $m_{ij\cdot 1}$.

It is now possible to eliminate the x_2 with the use of $m_{22\cdot 1}$ as a pivot. We first define

$$(4)\qquad m_{ij\cdot 12} = m_{22\cdot 1}m_{ij\cdot 1} - m_{i2\cdot 1}m_{2j\cdot 1}$$

and eliminate the x_2 from (3). We get

$$\begin{aligned} m_{33\cdot 12}x_3 + m_{34\cdot 12}x_4 &= m_{35\cdot 12} \\ m_{43\cdot 12}x_3 + m_{44\cdot 12}x_4 &= m_{45\cdot 12}. \end{aligned} \tag{5}$$

The elimination of x_3, using $m_{33\cdot 12}$ as a pivot, results in

$$m_{44\cdot 123}x_4 = m_{45\cdot 123} \tag{6}$$

with

$$m_{ij\cdot 123} = m_{33\cdot 12}m_{ij\cdot 12} - m_{i3\cdot 12}m_{3j\cdot 12} \tag{7}$$

so that

$$x_4 = \frac{m_{45\cdot 123}}{m_{44\cdot 123}}. \tag{8}$$

The important feature of this pivotal condensation method is the successive calculation with the use of the operational unit $U_2 = ab - cd$. This operation is easily carried out with a modern computing machine, and it is also accomplished readily without machines if the coefficients are small integers. A problem of this latter type is used to illustrate the method so that large numbers do not distract from our understanding the simplicity of the scheme. We wish to find the value of x_i in the system of equations

$$\begin{aligned} 2x_1 - x_2 + x_3 - x_4 &= 1 \\ -x_1 + 2x_2 - x_3 + x_4 &= 1 \\ x_1 - x_2 + 2x_3 - x_4 &= 1 \\ -x_1 + x_2 - x_3 + 2x_4 &= 1. \end{aligned} \tag{9}$$

The synthetic solution is shown in Table 4.2*a*. It is not necessary to indicate the symbols x_1, x_2, x_3, x_4 throughout the solution, since these are indicated, once and for all, by the columnar headings. The equals sign also need not be written since it is represented in the synthetic form by the vertical line at the right of the fourth column. The general presentation is given on the left and the application to (9) on the right of Table 4.2*a*. The check column and bottom row are explained in later sections.

Each $ab - cd$ operation can be described as follows: Multiply any entry by the leading element and subtract the product of the element at the left of its row and by the element at the top of its column. For example, $m_{35\cdot 12} = (3)(1) - (-1)(3) = 6$. This is computed directly from the table itself without reference to the formal definition

$$m_{35\cdot 12} = m_{22\cdot 1}m_{35\cdot 1} - m_{32\cdot 1}m_{25\cdot 1}.$$

The process is continued until the values $m_{44\cdot123} = 60$ and $m_{45\cdot123} = 60$ are obtained. From (8) it is clear that $x_4 = 1$. This value is placed in the appropriate position in the next-to-last row.

The equations obtained in the process of reducing the original equations may be called *elimination equations*, since they result from the elimination of one or more of the variables.

TABLE 4.2*a*

METHOD OF MULTIPLICATION AND SUBTRACTION

General						Illustration					
x_1	x_2	x_3	x_4		Check	x_1	x_2	x_3	x_4		Check
a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a_{1T}	2	−1	1	−1	1	2
a_{21}	a_{22}	a_{23}	a_{24}	a_{25}	a_{2T}	−1	2	−1	1	1	2
a_{31}	a_{32}	a_{33}	a_{34}	a_{35}	a_{3T}	1	−1	2	−1	1	2
a_{41}	a_{42}	a_{43}	a_{44}	a_{45}	a_{4T}	−1	1	−1	2	1	2
	$m_{22\cdot1}$	$m_{23\cdot1}$	$m_{24\cdot1}$	$m_{25\cdot1}$	$m_{2T\cdot1}$		3	−1	1	3	6
	$m_{32\cdot1}$	$m_{33\cdot1}$	$m_{34\cdot1}$	$m_{35\cdot1}$	$m_{3T\cdot1}$		−1	3	−1	1	2
	$m_{42\cdot1}$	$m_{43\cdot1}$	$m_{44\cdot1}$	$m_{45\cdot1}$	$m_{4T\cdot1}$		1	−1	3	3	6
		$m_{33\cdot12}$	$m_{34\cdot12}$	$m_{35\cdot12}$	$m_{3T\cdot12}$			8	−2	6	12
		$m_{43\cdot12}$	$m_{44\cdot12}$	$m_{45\cdot12}$	$m_{4T\cdot12}$			−2	8	6	12
			$m_{44\cdot123}$	$m_{45\cdot123}$	$m_{4T\cdot123}$				60	60	120
x_1	x_2	x_3	x_4			1	1	1	1		
$1 + x_1$	$1 + x_2$	$1 + x_3$	$1 + x_4$			2	2	2	2		

4.3 The back solution. Once the value of x_4 is obtained, it is possible to substitute in the preceding elimination equations to obtain x_3 and x_2, and in one of the original equations to obtain x_1. This process is known as the *back solution*. Thus, since

$$m_{33\cdot12}x_3 + m_{34\cdot12}x_4 = m_{35\cdot12}$$

(1)
$$m_{22\cdot1}x_2 + m_{23\cdot1}x_3 + m_{24\cdot1}x_4 = m_{25\cdot1}$$

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + a_{14}x_4 = a_{15}$$

we have

$$x_3 = \frac{(m_{35\cdot 12} - m_{34\cdot 12}x_4)}{m_{33\cdot 12}}$$

(2)
$$x_2 = \frac{(m_{25\cdot 1} - m_{24\cdot 1}x_4 - m_{23\cdot 1}x_3)}{m_{22\cdot 1}}$$

$$x_1 = \frac{(a_{15} - a_{14}x_4 - a_{13}x_3 - a_{12}x_2)}{a_{11}}.$$

Each of these operations is an operational unit (U_8 or U_9). The values needed for substitution in (2) are so arranged in Table 4.2*a* that the desired products are easily formed. Thus

$$x_3 = \frac{6 - (-2)(1)}{8} = 1$$

$$x_2 = \frac{3 - (1)(1) - (-1)(1)}{3} = 1$$

$$x_1 = \frac{1 - (-1)(1) - (1)(1) - (-1)(1)}{2} = 1.$$

Although the forward solution yields exact digital numbers, it is not to be expected that the back solution will necessarily yield integers (as in Table 4.2*a*) since divisions are involved. The back solution may be carried out exactly, using fractional forms, or a decimal approximation may be used. An illustration is given showing the exact fractional answers and the approximate decimal answers resulting from the two different types of back solution as applied to the equations

(3)
$$x + y - z = 7$$
$$x + 2y - 3z = 2$$
$$2x - y - 2z = 30.$$

The synthetic solution is presented in Table 4.3*a*, where fractional (exact) answers appear in the next-to-last row and decimal (approximate) answers in the last row:

TABLE 4.3*a*

FRACTIONAL AND DECIMAL ANSWERS

x	y	z		Check
1	1	−1	7	8
1	2	−3	2	2
2	−1	−2	30	29
	1	−2	−5	−6
	−3	0	16	13
		−6	1	−5
$\frac{73}{6}$	$\frac{-16}{3}$	$\frac{-1}{6}$		
12.1667	−5.3334	−0.1667		

4.4 The case with leading element zero. This procedure works out very well except when some one of the leading elements a_{11}, $m_{22\cdot 1}$, $m_{33\cdot 12}$, $\cdots$ is zero. It is not very satisfactory in this case, except when $m_{pp\cdot 12\cdots p-1}$ is zero, since the use of a zero pivot eliminates the linear independency of the equations by substituting equations that are linearly dependent. Thus the procedure applied to

$$\begin{aligned} 0x + a_{12}y + a_{13}z &= a_{14} \\ a_{21}x + a_{22}y + a_{23}z &= a_{24} \\ a_{31}x + a_{32}y + a_{33}z &= a_{34} \end{aligned} \tag{1}$$

yields at once

$$\begin{aligned} -a_{21}a_{12}y - a_{21}a_{13}z &= -a_{21}a_{14} \\ -a_{31}a_{12}y - a_{31}a_{13}z &= -a_{31}a_{14} \end{aligned} \tag{2}$$

and these equations are the first equation of (1) multiplied by $-a_{21}$ and $-a_{31}$, respectively. The contributions of the second and third equations of (1) have been annihilated by multiplication by zero so that it is impossible to get the solutions of (1) and (2) without using the second and third equations of (1) again.

It is well known that division by zero is excluded from algebraic and arithmetic operations. It is here indicated that multiplication by zero,

though permissible, is also to be avoided. Various adjustments are feasible when the leading term is zero. We can (*a*) interchange the equations so that the leading term is not zero, (*b*) interchange the order of the variables within the equations, (*c*) use the term directly under the leading term as a pivot, or (*d*) use a more general elimination method in which any element may be used as a pivot, as described in section 4.7.

TABLE 4.4*a*

INTERCHANGE OF VARIABLES

y	z	x	
1	−1	0	3
−1	−2	2	−3
1	1	1	−4
	−3	2	0
	2	1	−7
		−7	21
1	−2	−3	

TABLE 4.4*b*

INTERCHANGE OF EQUATIONS

x	y	z	
2	−1	−2	−3
0	1	−1	3
1	1	1	−4
	2	−2	6
	3	4	−5
		14	−28
−3	1	−2	

Tables 4.4*a* and 4.4*b* show how the first two of these methods may be used to solve the equations

$$\begin{aligned} 0x + y - z &= 3 \\ 2x - y - 2z &= -3 \\ x + y + z &= -4. \end{aligned} \tag{3}$$

4.5 Many variables. The details of the solution of (4.1.1) have been worked out with $p = 4$. The method can be applied in solving p equations in p unknowns by eliminating one variable at a time until a single equation is reached. The general term in the process is defined by

$$m_{ij\cdot 12\cdots h} = m_{hh\cdot 12\cdots h-1} m_{ij\cdot 12\cdots h-1} - m_{ih\cdot 12\cdots h-1} m_{hj\cdot 12\cdots h-1}. \tag{1}$$

The forward solution will finally terminate in the equation

$$m_{pp\cdot 12\cdots p-1} x_p = m_{p,\,p+1\cdot 12\cdots p-1} \tag{2}$$

from which x_p is obtained by division if $m_{pp\cdot 12\cdots p-1} \neq 0$. If the conventional order of elimination is used, the secondary subscripts are in

increasing order, and we need to write only the last one to indicate that all integers to $p - 1$ appear. Thus (2) can be written

$$m_{pp\cdot(p-1)}x_p = m_{p,\,p+1\cdot(p-1)}. \tag{3}$$

The back solution is carried out in the general case just as in the special case. Any of the elimination equations may be chosen for back-solution work, but it is inadvisable to choose one whose leading coefficient is zero.

4.6 Checking devices. The sum of the coefficients of a row may be used in eliminating "mistakes" and in studying the accumulation of errors. In the forward solution it is necessary only to record the sum of the entries for each row and to treat it like the other elements of that row. The calculated check value should be the sum of the entries of the row. A check column is used in Table 4.2*a* and Table 4.3*a*. Thus in Table 4.2*a*

$$8 + (-2) + 6 = 12$$

is the value of

$$m_{33\cdot 12} + m_{34\cdot 12} + m_{35\cdot 12} = m_{3T\cdot 12}.$$

This row sum check always holds. To show this, consider the elimination from the a_{ij} to the $m_{ij\cdot 1}$ since, if the property holds for this elimination, it will hold for the next elimination. Let

$$a_{iT} = a_{i1} + a_{i2} + a_{i3} + \cdots + a_{i,\,p+1}. \tag{1}$$

Then

$$\begin{aligned} m_{iT\cdot 1} = a_{11}a_{iT} - a_{i1}a_{1T} &= a_{11}(a_{i1} + a_{i2} + \cdots + a_{i,\,p+1}) \\ &\quad - a_{i1}(a_{11} + a_{12} + \cdots + a_{1,\,p+1}) \\ &= m_{i2\cdot 1} + m_{i3\cdot 1} + \cdots + m_{i,\,p+1\cdot 1}. \end{aligned} \tag{2}$$

This checking method is continued through the forward solution and can be carried through the back solution, as is illustrated in Table 4.2*a*. The values of x_i and the values of x_i obtained with the use of the row sum check column differ by unity.

The basic check that should always be made is a verification that the proposed solutions do actually satisfy all the original equations. The use of this final verification is advised as a general computational procedure. Thus in Table 4.2*a* the values of x_1, x_2, x_3, and x_4 are finally checked by multiplying each a_{ij} by the value of x_i below and adding to get the values a_{15}, a_{25}, a_{35}, and a_{45}. If this is done, we need not carry the check column through the back solution.

A similar column sum check can be applied.

We may prefer not to use a row sum check if few variables are involved. It has not been used in Tables 4.4*a* and 4.4*b*. If many variables are involved, it may be wise to insert additional check columns at regular intervals (say after each ten variables) to keep mistakes under control. In all cases, however, a final verification should be used.

4.7 Order of elimination. The equations may be arranged in such a way that any element may serve as a pivot. For example, the equations 4.1.2 may be written

$$
\begin{aligned}
a_{23}x_3 + a_{24}x_4 + a_{21}x_1 + a_{22}x_2 &= a_{25} \\
a_{33}x_3 + a_{34}x_4 + a_{31}x_1 + a_{32}x_2 &= a_{35} \\
a_{13}x_3 + a_{14}x_4 + a_{11}x_1 + a_{12}x_2 &= a_{15} \\
a_{43}x_3 + a_{44}x_4 + a_{41}x_1 + a_{42}x_2 &= a_{45}.
\end{aligned} \tag{1}
$$

This solution proceeds as in Table 4.2*a*. The details of the solution of (1) are presented in Table 4.7*a*.

TABLE 4.7*a*

ALTERNATIVE SOLUTION TO THAT OF TABLE 4.2*a*

x_3	x_4	x_1	x_2		Check
−1	1	−1	2	1	2
2	−1	1	−1	1	2
1	−1	2	−1	1	2
−1	2	−1	1	1	2
	−1	1	−3	−3	−6
	0	−1	−1	−2	−4
	−1	0	1	0	0
		1	1	2	4
		1	−4	−3	−6
			−5	−5	−10
1	1	1	1		

It is possible to perform the eliminations of Table 4.7*a* without rearranging the equations in the form (1) since any element may be used as a pivot. The technique is readily accomplished if the proposed pivotal element and the row and column in which it is located are

marked in some way. This may be done by drawing lines about the row and the column or by drawing a light line (perhaps of different color) through each. The $ab - cd$ technique continues as before, each element being multiplied by the pivotal element and the product elements being taken from the marked row and column. The operations of Table 4.7*a* are shown in Table 4.7*b*, which solves (4.1.2) with the use

TABLE 4.7*b*

USE OF NON-DIAGONAL PIVOTS

x_1	x_2	x_3	x_4		Check
2	−1	1	−1	1	2
−1	2	[−1]	1	1	2
1	−1	2	−1	1	2
−1	1	−1	2	1	2
−1	−1		0	−2	−4
1	−3		[−1]	−3	−6
0	1		−1	0	0
[1]	1			2	4
1	−4			−3	−6
	−5			−5	−10
1	1	1	1		

of non-diagonal pivots. The pivots that are successively -1, -1, and 1 are marked.

General formulas could be written for the case of the non-diagonal pivot, but it does not seem wise to take space to do this since the columns and rows may be rearranged so that the solution takes the usual form with diagonal pivot.

As was indicated in section 4.4, a non-diagonal pivot may be used when the diagonal term is zero without rearranging the equations.

We should consider the size of the coefficients of the different equations in designing the computational procedure. In general it is wise to use the equation having the smallest coefficients as the first equation (if there is any real difference in the size of the coefficients). It is also wise to use as the first variable the one having its coefficients smaller.

These principles of computational design would lead us to rearrange the equations

$$\begin{aligned} 100x_1 + 10x_2 + x_3 &= 111 \\ 25x_1 + 5x_2 + x_3 &= 31 \\ x_1 + x_2 + x_3 &= 3 \end{aligned} \tag{2}$$

in the form

$$\begin{aligned} x_3 + x_2 + x_1 &= 3 \\ x_3 + 5x_2 + 25x_1 &= 31 \\ x_3 + 10x_2 + 100x_1 &= 111 \end{aligned} \tag{3}$$

before the solution if we wish to feature small numbers. The solutions of (2) and (3) are presented in Table 4.7*c* for comparison.

TABLE 4.7*c*

REARRANGEMENT OF EQUATIONS

x_1	x_2	x_3	
100	10	1	111
25	5	1	31
1	1	1	3
	250	75	325
	90	99	189
		18,000	18,000
1	1	1	

x_3	x_2	x_1	
1	1	1	3
1	5	25	31
1	10	100	111
	4	24	28
	9	99	108
		180	180
1	1	1	

The arrangement (3) is preferable to (2). Of course the earlier methods of this section could be used to accomplish the same purpose without rearrangement. No hard and fast rule need be drawn up, but sound computational design calls for the use of small numbers whenever possible.

In most problems the coefficients are of the same order, so that no adjustment of this sort is wise. In any case, we usually do not wish to lose symmetry, if originally present, by some arbitrary rearrangement.

4.8 Use of symmetry. In many problems (particularly in those problems resulting from the application of least squares or regression theory) the coefficients of the unknowns have the symmetric property $a_{ij} = a_{ji}$. Matrices having this property are commonly called axisym-

metric or symmetric. The use of a diagonal pivot then insures the symmetry of the new set of equations. If

$$a_{ij} = a_{ji}, \qquad a_{ik} = a_{ki}, \qquad a_{jk} = a_{kj} \tag{1}$$

then

$$m_{ij\cdot k} = a_{kk}a_{ij} - a_{ik}a_{kj} = a_{kk}a_{ji} - a_{jk}a_{ki} = m_{ji\cdot k}. \tag{2}$$

By an extension of the same reasoning it can be shown that

$$m_{ij\cdot(h)} = m_{ji\cdot(h)}. \tag{3}$$

This fact may serve as the basis of a computational procedure in which all elements below the diagonal are omitted at each step. Any value of a_{ij} (or $m_{ij\cdot(h)}$) with $i > j$ can be found by using the corresponding values a_{ji} (or $m_{ji\cdot(h)}$). The general term (4.5.1) becomes

$$m_{ji\cdot(h)} = m_{ij\cdot(h)} = m_{hh\cdot(h-1)}m_{ij\cdot(h-1)} - m_{hi\cdot(h-1)}m_{hj\cdot(h-1)}$$

and the non-pivotal values are all in the ith and the jth columns.

Table 4.8*a* illustrates the method of multiplication and subtraction, using symmetry as applied to a numerical problem previously used [B].

The back solution is carried to four decimal places. Verification shows that these four decimal places are exact solutions of a set of equa-

TABLE 4.8*a*

METHOD OF MULTIPLICATION AND SUBTRACTION WITH SYMMETRY

x_1	x_2	x_3	x_4	a_{i5}	Check	a'_{i5}
1.0000	0.4000	0.5000	0.6000	0.2000	2.7000	0.20003
.4 500	1.0000	0.3000	0.4000	0.4000	2.5000	0.40001
.5000	.3 80	1.0000	0.2000	0.6000	2.6000	0.60001
.6000	*	*	1.0000	0.8000	3.0000	0.80004
	.2400	.1000				
	0.8400	0.1000	0.1600	0.3200	1.4200	
	*	0.7500	−0.1000	0.5000	1.2500	
	*	*	0.6400	0.6800	1.3800	
		0.6200	−0.1000	0.3880	0.9080	
		*	0.5120	0.5200	0.9320	
			0.30744	0.36120	0.66864	
−0.9366	0.0601	0.8153	1.1749			
0.0634	1.0601	1.8153	2.1749			

tions having the same coefficients on the left, but a slightly different set, a'_{i5} on the right. The a'_{i5} are recorded in Table 4.8*a*. Comparison of a'_{i5} with a_{i5} reveals that, in no case, does an a'_{i5} differ from the corresponding a_{i5} by more than 0.00004.

When the forward solution is exact it is relatively easy to find the approximate values of the x's to machine capacity if this is desired. The back solution of Table 4.8*a*, carried to nine decimal places, yields

$$\begin{aligned} x_1 &= -0.936612022 \\ x_2 &= 0.060109290 \\ x_3 &= 0.815300546 \\ x_4 &= 1.174863388. \end{aligned}$$

Comparison of these values with those of Table 4.8*a* shows that the values of Table 4.8*a* are correct to four decimal places, that is, they are the values obtained by taking a solution that is correct to a large number of places and rounding off to four. The deletion of entries below the diagonal is indicated by an asterisk. The basic computational procedure now calls for the multiplication of a given element a_{ij} by the leading element and the subtraction of the product of the elements heading column i and column j. Thus

$$\begin{aligned} m_{ij\cdot 1} &= a_{11}a_{ij} - a_{i1}a_{1j} \\ &= a_{11}a_{ij} - a_{1i}a_{1j}. \end{aligned} \tag{4}$$

For example, in Table 4.8*a*,

$$\begin{aligned} m_{34\cdot 1} &= a_{11}a_{34} - a_{13}a_{14} \\ &= (1.0000)(0.2000) - (0.5000)(0.6000) = -0.1000. \end{aligned}$$

4.9 Abbreviated methods. Many of the terms $m_{ij\cdot 1}$, $m_{ij\cdot 12}$, $m_{ij\cdot 123}$, $\cdots$ need not appear in the solution since it is essential that the first column and the first row only be recorded at each elimination. Methods resulting from the elimination of this recording are called *abbreviated methods*. Abbreviated methods make possible (*a*) a saving in time necessary for recording, (*b*) a saving in space since the calculational form may be designed more compactly, and (*c*) fewer mistakes since there is less recording and the numbers to be used are more accessible. The values to be recorded are computed as before with the use of the formulas

$$\begin{aligned} m_{ij\cdot 1} &= a_{11}a_{ij} - a_{1i}a_{1j} \\ m_{ij\cdot 12} &= m_{22\cdot 1}m_{ij\cdot 1} - m_{i2\cdot 1}m_{2j\cdot 1} \\ m_{ij\cdot 123} &= m_{33\cdot 12}m_{ij\cdot 12} - m_{i3\cdot 12}m_{3j\cdot 12}. \end{aligned}$$

If we compute the first row and the first column only at each step, we find that all the values necessary for continued multiplication and subtraction are available. To get the values $m_{ij\cdot 12}$, we use

$$(1) \qquad m_{ij\cdot 12} = m_{22\cdot 1}(a_{11}a_{ij} - a_{i1}a_{1j}) - m_{i2\cdot 1}m_{2j\cdot 1}$$

and the values of $m_{ij\cdot 123}$ are found with the use of

$$(2) \quad m_{ij\cdot 123} = m_{33\cdot 12}\,[m_{22\cdot 1}(a_{11}a_{ij} - a_{i1}a_{1j}) - m_{i2\cdot 1}m_{2j\cdot 1}] - m_{i3\cdot 12}m_{3j\cdot 12}.$$

Start with the a_{ij} and perform the indicated operations. Formulas (1) and (2) require the use of recording units rather than operational units since it is necessary to transfer the result at each stage, with most machines, from the products register to the setting mechanism. However, this recording unit is not unsatisfactory with machines currently available.

The abbreviated form of the method of Table 4.8*a* (not using symmetry) is presented in Table 4.9*a*. The appropriate terms of Table 4.9*a*

TABLE 4.9*a*

ABBREVIATED METHOD OF MULTIPLICATION AND SUBTRACTION

x_1	x_2	x_3	x_4	a_{i5}	Check	a'_{i5}
1.0000	0.4000	0.5000	0.6000	0.2000	2.7000	0.20003
0.4000	1.0000	0.3000	0.4000	0.4000	2.5000	0.40001
0.5000	0.3000	1.0000	0.2000	0.6000	2.6000	0.60001
0.6000	0.4000	0.2000	1.0000	0.8000	3.0000	0.80004
	0.8400	0.1000	0.1600	0.3200	1.4200	
	0.1000	*	*	*	*	
	0.1600	*	*	*	*	
		0.6200	−0.1000	0.3880	0.9080	
		−0.1000	*	*	*	
			0.30744	0.36120	0.66864	
−0.9366	0.0601	0.8153	1.1749			
0.0634	1.0601	1.8153	2.1749			

are easily found. Thus

$$x_{34\cdot 12} = 0.8400[(1.0000)(0.2000) - (0.5000)(0.6000)] - (0.1000)(0.1600)$$
$$= (0.8400)(-0.1000) - (0.1000)(0.1600)$$
$$= -0.1000.$$

The abbreviated form of Table 4.9*a* can be presented more compactly, as in Table 4.9*b*. The elements of row 10 of Table 4.9*a* may be

TABLE 4.9*b*

COMPACT METHOD OF MULTIPLICATION AND SUBTRACTION

x_1	x_2	x_3	x_4	a_{i5}	Check	a'_{i5}
1.0000	0.4000	0.5000	0.6000	0.2000	2.7000	0.20003
0.4000	1.0000	0.3000	0.4000	0.4000	2.5000	0.40001
0.5000	0.3000	1.0000	0.2000	0.6000	2.6000	0.60001
0.6000	0.4000	0.2000	1.0000	0.8000	3.0000	0.80004
1.0000	0.4000	0.5000	0.6000	0.2000	2.7000	
0.4000	0.8400	0.1000	0.1600	0.3200	1.4200	
0.5000	0.1000	0.6200	−0.1000	0.3880	0.9080	
0.6000	0.1600	−0.1000	0.30744	0.36120	0.66864	
−0.9366	0.0601	0.8153	1.1749			
0.0634	1.0601	1.8153	2.1749			

recorded in the ninth row, and the elements of the new eighth and ninth rows may be recorded in the sixth and seventh rows. A more satisfactory computational procedure results if the first row and the first column are written over again in the second half of the form.

Not only is Table 4.9*b* compact, but the arrangement is such that it is easy to find the elements necessary for calculation. With the exception of a_{ij}, all the values m needed for calculation are found in the second four rows. The $m_{ii\cdot(h)}$ elements appear in the diagonal, and the elements whose product is to be subtracted appear in converging row and column (ith row and jth column). Thus to compute $m_{34\cdot 12}$ in Table 4.9*b*, take $a_{34} = 0.2000$ and consider it placed in row 3 and column 4 of the second matrix. Multiply by 1.0000 (the first diagonal

term) and subtract the product (0.5000)(0.6000). The result is −0.1000. Take this result and multiply by 0.8400 (the second diagonal term) and subtract the product of the second entries in row 3 and column 4. Thus

$$m_{34\cdot 12} = (0.8400)(-0.1000) - (0.1000)(0.1600) = -0.1000.$$

In computing the row sum check, and in carrying out the back solution, we should not use entries to the left of the diagonal as the other entries are not a part of the elimination equations. In Table 4.9*b* vertical and horizontal lines are drawn to the left of the diagonal term to outline the elimination equations. These lines need not be inserted if we understand the meaning of each element in the compact solution.

4.10 Symmetric and abbreviated methods. If the equations are symmetric, we may combine the abbreviations of the last two sections. The results are similar to those of Table 4.9*b* except that no entries need be made to the left of the diagonal since $m_{ij\cdot(k)} = m_{ji\cdot(k)}$. The compact symmetric method of multiplication and subtraction is illustrated in Table 4.10*a*.

TABLE 4.10*a*

COMPACT SYMMETRIC METHOD OF MULTIPLICATION AND SUBTRACTION

x_1	x_2	x_3	x_4	a_{i5}	Sum	a'_{i5}
1.0000	0.4000	0.5000	0.6000	0.2000	2.7000	0.20003
*	1.0000	0.3000	0.4000	0.4000	2.5000	0.40001
*	*	1.0000	0.2000	0.6000	2.6000	0.60001
*	*	*	1.0000	0.8000	3.0000	0.80004
1.0000	0.4000	0.5000	0.6000	0.2000	2.7000	
	0.8400	0.1000	0.1600	0.3200	1.4200	
		0.6200	−0.1000	0.3880	0.9080	
			0.30744	0.36120	0.66864	
−0.9366	0.0601	0.8153	1.1749			
0.0634	1.0601	1.8153	2.1749			

The computational procedure is identical with that of the compact method with the exception that column i is now substituted for row i. The elements used in forming the subtracted products are found in column i and column j. Thus the elements 0.5000, 0.6000, 0.1000, 0.1600, used in computing $m_{34\cdot 1}$ and $m_{34\cdot 12}$ are found in the first two rows of the elimination matrix in columns 3 and 4.

4.11 Determinate, inconsistent, and equivalent equations. It is conventional to treat the question of determinateness, consistency, and equivalence of equations by means of determinants. However, determinants need not be introduced as the techniques and results of the method of multiplication and subtraction are immediately applicable.

Simultaneous equations that have one and only one (exact) solution are said to be *determinate*. All the simultaneous equations used thus far in this chapter are determinate.

Simultaneous equations that have infinitely many solutions are said to be *indeterminate*. Thus the equations

$$\begin{aligned} x - y &= 2 \\ 2x - 2y &= 4 \end{aligned} \tag{1}$$

are indeterminate since they have as solutions $x = 2 + k$, $y = k$, where k can take on any value. These equations are also called *equivalent* since any values (x,y) satisfying the first equation, satisfy the second also. Thus the system

$$\begin{aligned} x + y - z &= 1 \\ 2x - y + 2z &= 2 \\ 3x + z &= 3 \end{aligned} \tag{2}$$

is equivalent since the third equation is the sum of the first two equations.

Simultaneous linear equations that have one or more solutions are said to be *consistent*. All the systems of equations mentioned so far in this chapter are consistent.

Simultaneous equations that have no solution are said to be *inconsistent*. These are also indeterminate.

The method of multiplication and subtraction enables us to determine, as the solution proceeds, whether the equations are:

(*a*) Determinate and consistent (unique solution).
(*b*) Indeterminate and consistent (infinite number of solutions).
(*c*) Indeterminate and inconsistent (no solution).

The forward solution should be carried out without using zero as a pivot. If the back solution can be carried out in a way not involving a division by zero, the solution is determinate.

If the forward solution shows a row with every element zero, the equations are indeterminate and consistent, for such a situation reveals that one of the original equations is a linear combination of the others, that is, it can be obtained by multiplying the other equations by certain

values and adding. The computational procedure indicates just which equations are involved. Thus the synthetic solution of (2) in Table 4.11*a*

TABLE 4.11*a*

SOLUTION OF INDETERMINATE AND CONSISTENT EQUATIONS

x	y	z	
1	1	−1	1
2	−1	2	2
3	0	1	3
	−3	4	0
	−3	4	0
		0	0
$1-\frac{1}{3}k$	$\frac{4}{3}k$	k	

shows that the third equation is a linear combination of the first two equations.

Although division by zero is excluded, the value 0/0 may have any value, for example, k. With $z = k$, the back solution gives

$$y = \frac{0 - 4k}{(-3)} = \frac{4k}{3}$$

$$x = \frac{[1 - (-1)(k) - (1)(\frac{4}{3}k)]}{1} = 1 - \frac{k}{3}.$$

Substitution of these values in (2) shows that the equations are satisfied for any finite value of k.

The method of multiplication and subtraction enables us not only to determine if the equations are indeterminate and consistent, but also to exhibit an analytic statement of the infinity of solutions. In this respect it has a decided advantage over (common) determinantal methods.

If the last row of the forward solution reduces to two zeros, replace the unknown by k and carry out the back solution. If the back solution can be carried out completely, the equations are indeterminate and consistent, and an infinity of solutions is indicated.

If every element of every row has been reduced to zero prior to the last elimination, the forward solution should stop. The back solution

then proceeds by assigning values to each variable having a zero recorded as a coefficient.* For example, to solve the (obviously) dependent equations

$$\begin{aligned} x + y + z &= 1 \\ 2x + 2y + 2z &= 2 \\ 4x + 4y + 4z &= 4 \end{aligned} \tag{3}$$

by a synthetic form of the method of multiplication and subtraction, the forward solution proceeds through one step only, after which z is replaced by k, y by l. The back solution then yields $x = 1 - k - l$. The synthetic solution is shown in Table 4.11*b*.

TABLE 4.11*b*

INCOMPLETE FORWARD SOLUTION—INDETERMINATE EQUATIONS

x	y	z	
1	1	1	1
2	2	2	2
4	4	4	4
	0	0	0
	0	0	0
$1-k-l$	l	k	

The case of no solution may be identified with the values of $k/0$, since a zero appears as the left entry in the final elimination equation (no pivot being zero) with a non-zero term on the right. Evaluation of $k/0$ when $k \neq 0$ is impossible, and these equations have no common solution. Thus the synthetic form of the solution of

$$\begin{aligned} 3x - 2y + z &= -4 \\ x - y + 2z &= -3 \\ 5x - 4y + 5z &= 2 \end{aligned}$$

in Table 4.11*c* shows that the equations are inconsistent since $z = -36/0$. This inconsistency is exhibited in the middle step, since $-y + 5z = -5$ and $-2y + 10z = 26$ cannot be true simultaneously.

* Other coefficients are zero, but they are not recorded in the synthetic form.

TABLE 4.11*c*

SOLUTION OF INCONSISTENT EQUATIONS

x	y	z	
3	−2	1	−4
1	−1	2	−3
5	−4	5	2
	−1	5	−5
	−2	10	26
		0	−36

To sum up, the rules are very simple for determining the precise nature of the equations from the analytic form of the solution with the method of multiplication and subtraction. They are:

(*a*) Proceed with the forward solution, but do not use zero as a pivot.

(*b*) At any place in the back solution where the value $c/0$ (with $c \neq 0$) appears as one of the unknowns, it is indicated that the equations are inconsistent and that there is no solution.

(*c*) At any place in the back solution where the value $0/0$ appears as the value of one of the unknowns and where this indeterminancy cannot be resolved by using another equation, replace the value of $0/0$ by a parameter and proceed.

TABLE 4.11*d*

ADDITIONAL ILLUSTRATION

x	y	z	
1	1	1	1
1	1	−2	1
2	2	−1	2
	0	[−3]	0
	0	−3	0
	0		0
1 − k	k	0	

The problem of Table 4.11*d* serves as an illustration. Use the -3 as a pivot in line with the suggestions of sections 4.4 and 4.7. The forward solution then results in the value $y = 0/0 = k$. Then $z = (0 - 0 \cdot k)/(-3)$ and $x = 1 - k$.

4.12 Homogeneous equations. The general theory of the last section is immediately applicable to homogeneous equations in which the coefficients on the right side of (4.1.1) are zero. Homogeneous equations have the trivial solution $x = 0$, but the usual interest is in finding additional solutions if they exist. Since all the right terms are zero, the forward solution leads either to $0/0$, in which case the back solution is performed as before, or to the trivial case of $0/c$ when $c \neq 0$.

TABLE 4.12*a*

HOMOGENEOUS EQUATIONS—TRIVIAL SOLUTION

x	y	z	
1	2	3	0
3	−1	−2	0
1	1	−1	0
	−7	−11	0
	−1	−4	0
		17	0
0	0	0	

TABLE 4.12*b*

HOMOGENEOUS EQUATIONS WITH NON-TRIVIAL SOLUTION

x	y	z	
1	2	3	0
2	−1	2	0
5	0	7	0
	−5	−4	0
	−10	−8	0
		0	0
$-\frac{7}{5}k$	$-\frac{4}{5}k$	k	

An illustration with trivial solution only is presented in Table 4.12*a*.

If the last term on the left in the forward solution is not a zero, there is no solution, for if there were any other solution except the trivial solution, it would have to be obtained from an elimination equation whose leading element was zero. But there must be at least one equation at each step before the last in which the leading element is not zero if the last term on the left is not zero. It follows that the homogeneous equations have no solution except the trivial one if the left elimination entry does not equal zero.

A parameter may be introduced and the homogeneous equations may have an infinite number of solutions if the last elimination term on the left does equal zero. This situation is illustrated in Table 4.12*b*. Additional illustrations are presented in Tables 4.12*c* and 4.12*d*.

TABLE 4.12*c*

ADDITIONAL ILLUSTRATION USING HOMOGENEOUS EQUATIONS

x	y	z	
1	1	1	0
2	2	2	0
3	3	3	0
	0	0	0
	0	0	0
$-k-l$	l	k	

TABLE 4.12*d*

ADDITIONAL ILLUSTRATION USING HOMOGENEOUS EQUATIONS

x	y	z	
1	1	1	0
1	1	−2	0
2	2	−1	0
	0	−3	0
	0	−3	0
	0		0
$-k$	k	0	

4.13 Modification of the method of multiplication and subtraction. The earlier sections show that the method of multiplication and subtraction is very useful in obtaining exact solutions of simultaneous linear equations and in distinguishing the various types of solutions. The method, however, has two disadvantages:

(*a*) The continued multiplication may cause the numbers involved to become so large that they get beyond the capacity of the computing machine.

(*b*) With machines as they are now designed, the calculation of each $m_{ij\cdot(h)}$ is a recording unit rather than an operational unit. Operational units, of course, are to be preferred.

The following illustration shows how quickly problems involving two-place numbers can grow to machine capacity. The statement of the problem is found in the first four rows of the table and the solution of the problem in the last row. A row sum check is carried, and a final verification is made.

When the number of places becomes too large for machine capacity the method may be modified to give an approximate solution to a specified number of digits by dividing each entry in a given group of

TABLE 4.13*a*

METHOD OF MULTIPLICATION AND SUBTRACTION (LARGE NUMBERS)

x_1	x_2	x_3	x_4	a_{i5}	Sum
26	−10	15	32	23	86
19	45	−14	−8	57	99
−12	16	27	13	47	91
32	29	−35	28	−68	−14
	1360	−649	−816	1045	940
	296	882	722	1498	3398
	1074	−1390	−296	−2504	−3116
		1391624	1223456	1727960	4343040
		−1193374	473824	−4527770	−5247320
			2119425430720	−4238850861440	−2119425430720
2	1	3	−2		
3	2	4	−1		

equations by some appropriate power of ten and recording the rounded-off result. This method does not involve any great adjustment in the back solution if all terms in a given group are divided by the same power of ten. Thus if in Table 4.13*a* we divide the third group by 10^3 and round off, we have entries

1392	1223	1728	4343
−1193	474	−4528	−5247.

The complete (approximate) solution of the problem of Table 4.13*a*, using numbers of not more than four digits, is shown in Table 4.13*b*.

TABLE 4.13*b*

METHOD OF MULTIPLICATION AND SUBTRACTION (APPROXIMATE SOLUTION)

x_1	x_2	x_3	x_4	a_{i5}	Sum	a'_{i5}
26	−10	15	32	23	86	22.989
19	45	−14	−8	57	99	56.996
−12	16	27	13	47	91	46.932
32	29	−35	28	−68	−14	−67.990
	1360	−649	−816	1045	940	
	296	882	722	1498	3398	Check
	1074	−1390	−296	−2504	−3116	sum
10^3		1392	1223	1728	4343	4343
		−1193	474	−4528	−5247	−5247
10^9			2119	−4241	−2123	−2122
2.001	0.999	2.999	−2.001			

Comparison of the results of Table 4.13*b* with those of Table 4.13*a* show that no x_i, when recorded to three decimal places, is in error by more than 0.001.

Since there may now be some discrepancy between the row sum and the computed row sum, an additional column is added for exhibiting the actual row sum. The first four rows of this column may be used to indicate the values a'_{i5} obtained in the final verification. For purposes of verification it seems preferable to place this additional column before the computed check column, and this is done in the approximate methods appearing in the following chapters.

The x_i of Table 4.13*b* are the exact solutions of the equations in which the a'_{i5} replace the a_{i5}. The difference between these terms is in no case larger (in absolute value) than 0.07.

The powers of ten deleted at each stage of the elimination process may be placed at the left as shown in Table 4.13*b*.

REFERENCES

A. Descriptions of the method of multiplication and subtraction may be found in

1. P. S. Dwyer, "The solution of simultaneous equations," *Psychometrika*, **6**, 101–129 (1941). More general bibliography is presented here.
2. A. C. Aitken, *Determinants and Matrices*, second edition, Oliver and Boyd, Edinburgh, 1942.

B. This illustration has been used in the following papers to exhibit different methods, in addition to the first reference of A.

1. P. S. Dwyer, "The evaluation of determinants," *Psychometrika*, **6**, 191–204 (1941).
2. P. S. Dwyer, "The evaluation of linear forms," *Psychometrika*, **6**, 355–365 (1941).
3. P. S. Dwyer, "The Doolittle technique," *Annals of Mathematical Statistics*, **12**, 449–458 (1941).
4. Harold Hotelling, "Some new methods in matrix calculation," *Annals of Mathematical Statistics*, **14**, 1–34 (1943).

EXERCISES

1. Solve the equations

$$4x + 3y + 8z = 21$$
$$x - y - 2z = -3$$
$$3x + 4y - 7z = -4$$

by the method of multiplication and subtraction. Use a row sum check and verify the results.

2. Solve the equations

$$0x_1 - x_2 + 2x_3 = 18$$
$$2x_1 + 3x_2 + 4x_3 = 2$$
$$4x_1 - 3x_2 + x_3 = 41$$

(*a*) by rearranging the equations.
(*b*) with the use of a diagonal pivot.
(*c*) with the use of a non-diagonal pivot.

3. Obtain the answers to machine capacity of the symmetric system,

$$1.0x_1 + 0.4x_2 + 0.5x_3 = 0.2$$
$$0.4x_1 + 1.0x_2 + 0.3x_3 = 0.4$$
$$0.5x_1 + 0.3x_2 + 1.0x_3 = 0.6$$

with a non-abbreviated form of the method of multiplication and subtraction. Use a row sum check and verify the answers.

4. Work exercise 3, using a compact method.

In exercises 5 to 9, discuss the nature of the system of equations and provide solutions if they exist.

5.
$$x - 3y + 7z = 8$$
$$2x + 7y + 4z = 16$$
$$3x + 4y + 11z = 24.$$

6.
$$x - 3y + 7z = 8$$
$$2x - 6y + 14z = 16$$
$$3x + 4y + 11z = 24.$$

7.
$$x - 3y + 7z = 8$$
$$2x - 6y + 14z = 8$$
$$3x + 4y + 11z = 24.$$

8.
$$3x + 2y + z = 0$$
$$-2x - y + 3z = 0$$
$$-x + y + z = 0.$$

9.
$$3x + 2y + z = 0$$
$$2x - y + 2z = 0$$
$$7x + 0y + 5z = 0.$$

10. Obtain the approximate solution of

$$16x_1 - 19x_2 + 40x_3 - 17x_4 = 89$$
$$82x_1 + 42x_2 - 32x_3 - 8x_4 = 73$$
$$55x_1 - 35x_2 + 45x_3 - 25x_4 = 67$$
$$7x_1 + 19x_2 + 33x_3 + 85x_4 = 111$$

by the modified method of section 4.13. Use a row sum check and a final verification.

CHAPTER 5

The Method of Multiplication and Subtraction with (Exact) Division Method of Determinants

5.1 Introduction. As indicated in the preceding chapter the method of multiplication and subtraction may produce numbers that, though exact, contain many digits. This method may be modified in such a way as to feature (exact) numbers having a smaller number of digits by making use of the fact that the m's of Chapter 4 are themselves exactly divisible by certain pivots used in computing them. This feature is also very useful in checking. Such a modification could then be called the method of *multiplication and subtraction with (exact) division*. As is shown in section 10.4, every calculated term is a determinant of some of the coefficients of the equations, and hence we may call this a *method of determinants*.

A form of a method of multiplication and subtraction with exact division was introduced by Dodgson, who modified the previous work of Hermite and Chiò [A.1, 2, 3]. His method was a condensation method, but it did not use a fixed pivot. Aitken in 1932 [A.4] outlined a pivotal method of multiplication and subtraction with exact division that is similar to the method used in this chapter in solving simultaneous linear equations. Aitken did not emphasize the use of diagonal pivots. Waugh and Dwyer in 1945 [A.2] introduced a notation, similar to that of Chapter 4, that is especially applicable to the use of diagonal pivots, and they showed how the method could be used in forming a compact solution of linear equations and a compact calculation of the adjoint. The material used in this chapter is similar to that of the Waugh-Dwyer paper. Application is made to many of the illustrations of the previous chapter.

5.2 The forward solution. As in Chapter 4, we consider the solution of (4.2.1) and use (4.2.2) to illustrate the general theory.

The first elimination step in this solution is identical with the first step of the previous method, though a different notation is used. We define

$$d_{ij\cdot 1} = a_{11}a_{ij} - a_{i1}a_{1j}. \tag{1}$$

We eliminate the x_1 as before, and arrive at

$$\begin{aligned} d_{22\cdot 1}x_2 + d_{23\cdot 1}x_3 + d_{24\cdot 1}x_4 &= d_{25\cdot 1} \\ d_{32\cdot 1}x_2 + d_{33\cdot 1}x_3 + d_{34\cdot 1}x_4 &= d_{35\cdot 1} \\ d_{42\cdot 1}x_2 + d_{43\cdot 1}x_3 + d_{44\cdot 1}x_4 &= d_{45\cdot 1}. \end{aligned} \tag{2}$$

These equations are identical with (4.2.3) since $d_{ij\cdot 1} = m_{ij\cdot 1}$.

In considering the elimination of x_2 we note that each $m_{ij\cdot 12}$ of (4.2.4) and (4.2.5) is exactly divisible by a_{11}. This is shown by the fact that

$$\begin{aligned} m_{ij\cdot 12} &= m_{22\cdot 1}m_{ij\cdot 1} - m_{i2\cdot 1}m_{2j\cdot 1} \\ &= (a_{11}a_{22} - a_{21}a_{12})(a_{11}a_{ij} - a_{i1}a_{1j}) \\ &\quad - (a_{11}a_{i2} - a_{i1}a_{12})(a_{11}a_{2j} - a_{21}a_{1j}). \end{aligned} \tag{3}$$

When the right side of (3) is expanded it is evident that a_{11} is a factor of every term except the two product terms $a_{21}a_{12}a_{i1}a_{1j}$ and $-a_{i1}a_{12}\cdot a_{21}a_{1j}$. These terms cancel so that $m_{ij\cdot 12}$ is divisible by a_{11}. We then define

$$\begin{aligned} d_{ij\cdot 12} = \frac{m_{ij\cdot 12}}{a_{11}} &= \frac{m_{22\cdot 1}m_{ij\cdot 1} - m_{i2\cdot 1}m_{2j\cdot 1}}{a_{11}} \\ &= \frac{d_{22\cdot 1}d_{ij\cdot 1} - d_{i2\cdot 1}d_{2j\cdot 1}}{a_{11}}. \end{aligned} \tag{4}$$

This exact divisibility may be incorporated into the next elimination step since every $m_{ij\cdot 12}$ of (4.2.4) is divisible by a_{11}. As each $m_{ij\cdot 12}$ appears in the calculator, it is immediately divided by a_{11} and the result (which is $d_{ij\cdot 12}$) is recorded. If the division is not exact, a mistake has been made, and it should be found at once.

After x_2 is eliminated by computing the $d_{ij\cdot 12}$, we move on to the elimination of x_3. Now it can be shown, similarly, that $d_{22\cdot 1}$ divides $d_{33\cdot 12}d_{ij\cdot 12} - d_{i3\cdot 12}d_{3j\cdot 12}$ so that we may define

$$d_{ij\cdot 123} = \frac{(d_{33\cdot 12}d_{ij\cdot 12} - d_{i3\cdot 12}d_{3j\cdot 12})}{d_{22\cdot 1}}, \tag{5}$$

with the knowledge that the indicated division will yield a quotient with no remainder.

The same line of reasoning may be used to show that in general

$$\frac{d_{hh\cdot(h-1)}d_{ij\cdot(h-1)} - d_{ih\cdot(h-1)}d_{hj\cdot(h-1)}}{d_{h-1,h-1\cdot(h-2)}} = d_{ij\cdot(h)}, \tag{6}$$

where $d_{ij\cdot(h)} = d_{ij\cdot 12\cdots h}$ is an exact quotient.

This feature of exact divisibility is the basis of this modification of the method of multiplication and subtraction. It is similar to the method of Chapter 4, but utilizes the d's rather than the m's throughout. In addition to the exact divisibility feature that cuts down the number of digits and permits continuous checking, the method has theoretical advantages since each $d_{ij\cdot(h)}$ is identifiable as a determinant, as is shown in Chapter 10.

Table 5.2a shows the application of the method to the problem of

TABLE 5.2a

METHOD OF MULTIPLICATION AND SUBTRACTION WITH (EXACT) DIVISION

General						Illustration					
x_1	x_2	x_3	x_4		Sum	x_1	x_2	x_3	x_4		Sum
a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a_{1T}	2	−1	1	−1	1	2
a_{21}	a_{22}	a_{23}	a_{24}	a_{25}	a_{2T}	−1	2	−1	1	1	2
a_{31}	a_{32}	a_{33}	a_{34}	a_{35}	a_{3T}	1	−1	2	−1	1	2
a_{41}	a_{42}	a_{43}	a_{44}	a_{45}	a_{4T}	−1	1	−1	2	1	2
	$d_{22\cdot1}$	$d_{23\cdot1}$	$d_{24\cdot1}$	$d_{25\cdot1}$	$d_{2T\cdot1}$		3	−1	1	3	6
	$d_{32\cdot1}$	$d_{33\cdot1}$	$d_{34\cdot1}$	$d_{35\cdot1}$	$d_{3T\cdot1}$		−1	3	−1	1	2
	$d_{42\cdot1}$	$d_{43\cdot1}$	$d_{44\cdot1}$	$d_{45\cdot1}$	$d_{4T\cdot1}$		1	−1	3	3	6
		$d_{33\cdot12}$	$d_{34\cdot12}$	$d_{35\cdot12}$	$d_{3T\cdot12}$			4	−1	3	6
		$d_{43\cdot12}$	$d_{44\cdot12}$	$d_{45\cdot12}$	$d_{4T\cdot12}$			−1	4	3	6
			$d_{44\cdot123}$	$d_{45\cdot123}$	$d_{4T\cdot123}$				5	5	10
x_1	x_2	x_3	x_4			1	1	1	1		

Table 4.2a. The general presentation is given on the left and the illustration on the right.

The forward solution for the case $p = 4$ reduces to

$$x_4 = \frac{d_{45\cdot123}}{d_{44\cdot123}}. \tag{7}$$

The value of x_4 is recorded to some specified number of decimal places although an exact fractional form may be used if desired.

In like manner (4.2.1) leads to

$$x_p = \frac{d_{p,\,p+1\cdot(p-1)}}{d_{p,p\cdot(p-1)}}. \tag{8}$$

As with the method of multiplication and subtraction, no zero term should be used as a pivot.

5.3 The back solution. The back solution is carried out as in the last chapter with

$$\begin{aligned} x_4 &= \frac{d_{45\cdot123}}{d_{44\cdot123}} \\ x_3 &= \frac{(d_{35\cdot12} - d_{34\cdot12}x_4)}{d_{33\cdot12}} \\ x_2 &= \frac{(d_{25\cdot1} - d_{24\cdot1}x_4 - d_{23\cdot1}x_3)}{d_{22\cdot1}} \\ x_1 &= \frac{(a_{15} - a_{14}x_4 - a_{13}x_3 - a_{12}x_2)}{a_{11}}. \end{aligned} \tag{1}$$

The back solution is recorded in the bottom row of Table 5.2*a*. The exact divisibility feature of the forward solution does not necessarily carry to the back solution if decimals are used. It is wise to round off each quotient as it appears in the back solution to some specified number of decimal places, or, if one prefers, exact values may be obtained with the use of fractions. The first method is illustrated in Table 5.3*a* and the second in Table 5.3*b*. If the decimal approximation system is used, each step of (1) is an operational unit U_9.

TABLE 5.3*a*

ILLUSTRATION

x	y	z		Check
3	1	2	10	9.999
−1	2	3	8	8.000
2	−1	2	6	6.000
	7	11	34	
	−5	2	−2	
		23	52	
1.391	1.304	2.261		

TABLE 5.3*b*

ILLUSTRATION

x	y	z	
3	1	2	10
−1	2	3	8
2	−1	2	6
	7	11	34
	−5	2	−2
		23	52
$1\frac{9}{23}$	$1\frac{7}{23}$	$2\frac{6}{23}$	

5.4 Relations between the m's and the d's. The m's are multiples of the d's, and it is of interest to see what relations exist between them. Starting with the definitions $d_{ij\cdot(h)}$, we get successively

$$d_{ij\cdot 1} = m_{ij\cdot 1}$$

$$d_{ij\cdot 12} = \frac{(d_{22\cdot 1}d_{ij\cdot 1} - d_{i2\cdot 1}d_{2j\cdot 1})}{a_{11}} = \frac{m_{ij\cdot 12}}{a_{11}}$$

$$(1) \qquad d_{ij\cdot 123} = \frac{(d_{33\cdot 12}d_{ij\cdot 12} - d_{i3\cdot 12}d_{3j\cdot 12})}{d_{22\cdot 1}}$$

$$= \frac{1}{a_{11}^{\,2}}\left[\frac{m_{33\cdot 12}m_{ij\cdot 12} - m_{i3\cdot 12}m_{3j\cdot 12}}{m_{22\cdot 1}}\right]$$

$$= \frac{m_{ij\cdot 123}}{a_{11}^{\,2}m_{22\cdot 1}}.$$

In a similar manner

$$(2) \qquad d_{ij\cdot 1234} = \frac{m_{ij\cdot 1234}}{a_{11}^{\,3}m_{22\cdot 1}^{\,2}m_{33\cdot 12}}.$$

It can be proved by mathematical induction that

$$(3) \qquad d_{ij\cdot(h)} = \frac{m_{ij\cdot(h)}}{a_{11}^{\,h-1}m_{22\cdot 1}^{\,h-2}\cdots m_{h-1,h-1\cdot(h-2)}},$$

where no pivotal element is zero.

The formulas giving the values of the m's in terms of the d's are obtained by solving (1), (2), and (3) for the m's. Thus

$$m_{ij\cdot 1} = d_{ij\cdot 1}$$

$$m_{ij\cdot 12} = a_{11}d_{ij\cdot 12}$$

$$m_{ij\cdot 123} = a_{11}^{\,2}d_{22\cdot 1}d_{ij\cdot 123}$$

$$(4) \quad m_{ij\cdot 1234} = m_{ij\cdot(4)} = a_{11}^{\,4}d_{22\cdot 1}^{\,2}d_{33\cdot 12}d_{ij\cdot(4)}$$

$$m_{ij\cdot(5)} = a_{11}^{\,8}d_{22\cdot 1}^{\,4}d_{33\cdot 12}^{\,2}d_{44\cdot 123}d_{ij\cdot(5)}$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$m_{ij\cdot(h)} = (a_{11})^{(2^{h-2})}(d_{22\cdot 1})^{(2^{h-3})}(d_{33\cdot 12})^{(2^{h-4})}\cdots d_{h-1,h-1\cdot(h-2)}d_{ij\cdot(h)}.$$

5.5 The case with diagonal pivot zero. Since no zero element can be used as a pivot, a variation of the usual computational procedure must be made if the leading diagonal term is zero. It is suggested that

the methods of Chapter 4 be used with the additional division, by the previous pivot, added. The reader is referred to section 5.7 for further details.

5.6 Checking devices. The row sum check works just as it did in Chapter 4, since (4.6.2) now becomes

$$d_{iT\cdot 1} = d_{i2\cdot 1} + d_{i3\cdot 1} + \cdots + d_{i,\,p+1\cdot 1}. \tag{1}$$

At the next elimination the check equation is

$$d_{iT\cdot 12} = d_{i3\cdot 12} + d_{i4\cdot 12} + \cdots + d_{i,\,p+1\cdot 12}. \tag{2}$$

The last problem of Chapter 4 (Table 4.13a) is used to illustrate the check and to show the general advantage of the method. The solution is shown in Table 5.6.

TABLE 5.6a

ILLUSTRATION OF TABLE 4.13a

x_1	x_2	x_3	x_4		Sum
26	−10	15	32	23	86
19	45	−14	−8	57	99
−12	16	27	13	47	91
32	29	−35	28	−68	−14
	1360	−649	−816	1045	940
	296	882	722	1498	3398
	1074	−1390	−296	−2504	−3116
		53524	47056	66460	167040
		−45899	18224	−174145	−201820
			2305327	−4610654	−2305327
2	1	3	−2		

The method of multiplication and subtraction with (exact) division has all the checking features of the method of Chapter 4, and, in addition, each computed element after those of the first computed matrix is obtained as a result of an exact division. If this division is not exact, the calculated value is not correct. This check is of no value, of course, if the pivot is ± 1 or if fractions are used.

5.7 Order of elimination. Just as in Chapter 4 the equations may be rearranged in such a way that any element may serve as a pivot, or a non-diagonal pivot may be used. Illustrations corresponding to Tables 4.7*a* and 4.7*b* are presented in Tables 5.7*a* and 5.7*b*.

All the pivots used in Tables 5.7*a* and 5.7*b* are 1 or -1 so that, for this case, the solution of this chapter is no better than that previously

TABLE 5.7*a*

INTERCHANGE OF COLUMNS

x_3	x_4	x_1	x_2		Sum
−1	1	−1	2	1	2
2	−1	1	−1	1	2
1	−1	2	−1	1	2
−1	2	−1	1	1	2
	−1	1	−3	−3	−6
	0	−1	−1	−2	−4
	−1	0	1	0	0
		−1	−1	−2	−4
		−1	4	3	6
			5	5	10
1	1	1	1		

TABLE 5.7*b*

USE OF NON-DIAGONAL PIVOT

x_1	x_2	x_3	x_4		Sum
2	−1	1	−1	1	2
−1	2	[−1]	1	1	2
1	−1	2	−1	1	2
−1	1	−1	2	1	2
−1	−1		0	−2	−4
1	−3		[−1]	−3	−6
0	1		−1	0	0
[−1]	−1			−2	−4
−1	4			3	6
	5			5	10
1	1	1	1		

given. In general, use of 1 and -1 as pivots is to be avoided if the divisibility check is to be used.

Aitken's presentation [A.4] featured the use of non-diagonal pivots. Where non-diagonal pivots are used, it may be wise to eliminate the equation first whose coefficients are smaller than the coefficients of other equations. It may be wise also to use the numbers having small absolute values as pivots, although we should avoid $+1$, 0, and -1.

Usually we wish to use diagonal pivots, but occasionally one of these is zero. Suggested variations of the method not calling for rearrangement of the equations are:

(*a*) Use the element directly under the leading element as a pivot.
(*b*) Use the element to the right of the leading element as a pivot.
(*c*) Use the next diagonal element as a pivot.

TABLE 5.7c

LEADING DIAGONAL PIVOT ZERO 1

x	y	z	w	
2	1	1	−2	3
2	1	2	−1	5
1	−1	−1	−1	0
3	2	1	1	3
	0	2	2	4
	[−3]	−3	0	−3
	1	−1	8	−3
		−3	−3	−6
		3	−12	6
			−15	0
1	−1	2	0	

LEADING DIAGONAL PIVOT ZERO 2

x	y	z	w	
2	1	1	−2	3
2	1	2	−1	5
1	−1	−1	−1	0
3	2	1	1	3
	0	[2]	2	4
	−3	−3	0	−3
	1	−1	8	−3
	[−3]		3	3
	1		9	−1
			−15	0
1	−1	2	0	

LEADING DIAGONAL PIVOT ZERO 3

x	y	z	w	
2	1	1	−2	3
2	1	2	−1	5
1	−1	−1	−1	0
3	2	1	1	3
	0	2	2	4
	−3	[−3]	0	−3
	1	−1	8	−3
	[3]		−3	−3
	−3		−12	3
			15	0
1	−1	2	0	

These methods are illustrated in Table 5.7*c*, where the equations

$$\begin{aligned} 2x + y + z - 2w &= 3 \\ 2x + y + 2z - w &= 5 \\ x - y - z - w &= 0 \\ 3x + 2y + z + w &= 3 \end{aligned} \tag{1}$$

are solved by these methods.

The first of these methods is probably the easiest, but the third method maintains the symmetry of the matrix of the coefficients if symmetry is originally present.

In connection with the order of elimination, we naturally question the relation between $d_{ii\cdot j}$ and $d_{jj\cdot i}$, $d_{ij\cdot kl}$ and $d_{ji\cdot kl}$, etc. Now

$$d_{ii\cdot j} = d_{jj\cdot i} \tag{2}$$

since

$$d_{ii\cdot j} = a_{jj}a_{ii} - a_{ji}a_{ij} \qquad \text{and} \qquad d_{jj\cdot i} = a_{ii}a_{jj} - a_{ij}a_{ji}.$$

Similarly, using (5.2.6), we obtain

$$\begin{aligned} d_{ij\cdot kl} &= \frac{(d_{ll\cdot k}d_{ij\cdot k} - d_{il\cdot k}d_{lj\cdot k})}{a_{kk}} \\ &= a_{kk}a_{ll}a_{ij} - (a_{lk}a_{kl}a_{ij}) - (a_{ll}a_{ik}a_{kj} + a_{il}a_{kk}a_{lj}) \\ &\quad + (a_{il}a_{lk}a_{kj} + a_{ik}a_{kl}a_{lj}). \end{aligned} \tag{3}$$

Now (3) features four terms, each of which is symmetric in k and l, that is, a substitution of k for l and l for k does not change the term. Hence

$$d_{ij\cdot kl} = d_{ij\cdot lk}. \tag{4}$$

If we let the values of d_{ij} be the values of a_{ij}, we can use (4) to show that

$$d_{ij\cdot klm} = d_{ij\cdot kml}. \tag{5}$$

It follows from the use of (4), which permits the interchange of the first and second secondary subscripts, and from (5), which permits the interchange of the second and third secondary subscripts, that

$$d_{ij\cdot klm} = d_{ji\cdot kml} = d_{ij\cdot mkl} = d_{ji\cdot lkm} = d_{ij\cdot lkm}. \tag{6}$$

This argument can be extended to any number of secondary subscripts. The numerical result is independent of the order of elimination.

The special case $i = j$ is interesting for it indicates a diagonal term. Since $d_{ii\cdot j} = d_{jj\cdot i}$ the primary subscripts may be interchanged with the secondary subscripts and

$$d_{ii\cdot jkl} = d_{jj\cdot ikl} = d_{kk\cdot ijl} = d_{\cdot ijkl}. \tag{7}$$

Further interpretation of these formulas is made in Chapter 10.

5.8 Symmetric methods. If the method of multiplication and subtraction with (exact) division is applied to a symmetric matrix and a diagonal term is used as a pivot, the resulting matrix is symmetric. In order to establish this we first consider the case in which the diagonal term is the leading element of the matrix. It is shown in Chapter 4 that $m_{ij\cdot(h)} = m_{ji\cdot(h)}$ if $a_{ij} = a_{ji}$. Application of (5.4.3) shows that $d_{ij\cdot(h)} = d_{ji\cdot(h)}$.

Next consider the matrix of the values $d_{ij\cdot 3214}$. The equations can be arranged with x_3 in the first position, x_2 in the second position, x_1 in the third position, and x_4 in the fourth position. If the original equations are also arranged in 3, 2, 1, 4 order, the resulting system is symmetric. The result of the last paragraph can be applied to show that the equations resulting from each elimination are symmetric if a diagonal pivot is used. A similar statement can be made for any other sequence of secondary subscripts.

As in section 4.8, the terms below the diagonal may be deleted. An illustration of this is shown in Table 5.8*b*. Table 5.8*a* shows the solu-

TABLE 5.8*a*

ILLUSTRATION OF TABLE 4.8*a*

x_1	x_2	x_3	x_4	a_{i5}	Sum	a'_{i5}
1.0000	0.4000	0.5000	0.6000	0.2000	2.7000	0.20003
0.4000	1.0000	0.3000	0.4000	0.4000	2.5000	0.40001
0.5000	0.3000	1.0000	0.2000	0.6000	2.6000	0.60001
0.6000	0.4000	0.2000	1.0000	0.8000	3.0000	0.80004
	0.8400	0.1000	0.1600	0.3200	1.4200	
	0.1000	0.7500	−0.1000	0.5000	1.2500	
	0.1600	−0.1000	0.6400	0.6800	1.3800	
		0.6200	−0.1000	0.3880	0.9080	
		−0.1000	0.5120	0.5200	0.9320	
			0.3660	0.4300	0.7960	
−0.9366	0.0601	0.8153	1.1749			

TABLE 5.8*b*

ILLUSTRATION OF TABLE 5.8*a*, USING SYMMETRY

x_1	x_2	x_3	x_4	a_{i5}	Sum	a'_{i5}
1.0000	0.4000	0.5000	0.6000	0.2000	2.7000	0.20003
*	1.0000	0.3000	0.4000	0.4000	2.5000	0.40001
*	*	1.0000	0.2000	0.6000	2.6000	0.60001
*	*	*	1.0000	0.8000	3.0000	0.80004
	0.8400	0.1000	0.1600	0.3200	1.4200	
	*	0.7500	−0.1000	0.5000	1.2500	
	*	*	0.6400	0.6800	1.3800	
		0.6200	−0.1000	0.3880	0.9080	
		*	0.5120	0.5200	0.9320	
			0.3660	0.4300	0.7960	
−0.9366	0.0601	0.8153	1.1749			

tion of Table 4.8*a* by the method of multiplication and subtraction with (exact) division without making use of symmetry.

5.9 Abbreviated methods. It is not necessary to write each $d_{ij\cdot 1}$, $d_{ij\cdot 12}$, $d_{ij\cdot 123}$, etc., but the first column and the first row of each matrix only. The solution of the problems of Tables 4.9*a* and 4.9*b*, using the d's, is indicated in Tables 5.9*a* and 5.9*b*.

Table 5.9*a* is a duplication of Table 5.8*a*, with some of the d's omitted as indicated above. Table 5.9*b* is a compact presentation of the work of Table 5.9*a*. As an illustration, the computation of $d_{45\cdot 123}$ is obtained as follows:

$$(0.8000)(1.0000) - (0.2000)(0.6000) = 0.6800$$

$$\frac{(0.6800)(0.8400) - (0.3200)(0.1600)}{1.0000} = 0.5200$$

$$\frac{(0.5200)(0.6200) - (0.3880)(-0.1000)}{0.8400} = 0.4300.$$

With present equipment the computation of a d may not be described as an operational unit since it is necessary to transfer the result of each

$ab - cd$ operation from the products register to the setting mechanism. A recent model Monroe, however, does permit, subject to certain restrictions, the automatic transfer of the result to the setting mechanism.

TABLE 5.9*a*

ABBREVIATED MULTIPLICATION AND SUBTRACTION WITH DIVISION

x_1	x_2	x_3	x_4	a_{i5}	Sum	a'_{i5}
1.0000	0.4000	0.5000	0.6000	0.2000	2.7000	0.20003
0.4000	1.0000	0.3000	0.4000	0.4000	2.5000	0.40001
0.5000	0.3000	1.0000	0.2000	0.6000	2.6000	0.60001
0.6000	0.4000	0.2000	1.0000	0.8000	3.0000	0.80004
	0.8400	0.1000	0.1600	0.3200	1.4200	
	0.1000	*	*	*	*	
	0.1600	*	*	*	*	
		0.6200	−0.1000	0.3880	0.9080	
		−0.1000	*	*	*	
			0.3660	0.4300	0.7960	
−0.9366	0.0601	0.8153	1.1749			

TABLE 5.9*b*

COMPACT MULTIPLICATION AND SUBTRACTION WITH DIVISION

x_1	x_2	x_3	x_4	a_{i5}	Sum	a'_{i5}
1.0000	0.4000	0.5000	0.6000	0.2000	2.7000	0.20003
0.4000	1.0000	0.3000	0.4000	0.4000	2.5000	0.40001
0.5000	0.3000	1.0000	0.2000	0.6000	2.6000	0.60001
0.6000	0.4000	0.2000	1.0000	0.8000	3.0000	0.80004
1.0000	0.4000	0.5000	0.6000	0.2000	2.7000	
0.4000	0.8400	0.1000	0.1600	0.3200	1.4200	
0.5000	0.1000	0.6200	−0.1000	0.3880	0.9080	
0.6000	0.1600	−0.1000	0.3660	0.4300	0.7960	
−0.9366	0.0601	0.8153	1.1749			

5.10 Abbreviated symmetric methods. The reader who has mastered the abbreviated symmetric methods of the last chapter should have no trouble in understanding how these methods are used in computing the d's. No entries below the main diagonal are recorded, and the product terms are taken from the top of the columns. Thus in getting $d_{45\cdot 123}$, as in the last section, the identical numbers are multiplied, but they are now found in parallel columns. This arrangement seems more satisfactory in problems involving many variables and is less conducive to mistakes. It also tends to make the back solution easier since no additional quantities are recorded in the row.

The indicated solution is presented in Table 5.10.

TABLE 5.10*a*

SYMMETRIC COMPACT MULTIPLICATION AND SUBTRACTION WITH DIVISION

x_1	x_2	x_3	x_4	a_{i5}	Sum	a'_{i5}
1.0000	0.4000	0.5000	0.6000	0.2000	2.7000	0.20003
*	1.0000	0.3000	0.4000	0.4000	2.5000	0.40001
*	*	1.0000	0.2000	0.6000	2.6000	0.60001
*	*	*	1.0000	0.8000	3.0000	0.80004
1.0000	0.4000	0.5000	0.6000	0.2000	2.7000	
	0.8400	0.1000	0.1600	0.3200	1.4200	
		0.6200	−0.1000	0.3880	0.9080	
			0.3660	0.4300	0.7960	
−0.9366	0.0601	0.8153	1.1749			

If a leading diagonal term is zero, we must adjust our procedure, just as in the last chapter, so as to use some term other than zero as a pivot. If the matrix is symmetrical, it is desirable that the symmetry be maintained, and for this purpose the use of some other diagonal pivot is indicated.

It is possible to use the same rules laid down in sections 4.11 and 4.12 regarding the nature of the solutions (and their explicit form) for both non-homogeneous and homogeneous equations when the forward solution has been carried out with the use of the d's.

5.11 Modification. It is possible to modify this method by dividing the equations at some step in the process by a power of 10 and then rounding off the result to some approximate answer. This is not as necessary as with the m's since the d's usually have a smaller number of

digits. Thus the d's of Table 5.6 are seven-place numbers or less, and rounding off is not advised as it was in Table 4.13a.

If we do approximate by rounding off, we lose the exact division feature, which is one of the most valuable advantages of the method. However, in a problem involving many variables, the d's may exceed the machine capacity so that it may be necessary to round off.

In concluding this chapter, it should be stated again that further identification of the d's is made in Chapter 10.

REFERENCES

A. Appropriate references may be found in

1. E. T. Whittaker and G. Robinson, *The Calculus of Observations*, Blackie and Son, Limited, London and Glasgow, fourth edition, 1944, pp. 71–77.
2. F. V. Waugh and P. S. Dwyer, "Compact computation of the inverse of a matrix," *Annals of Mathematical Statistics*, **16**, 259–271 (1945).
3. C. L. Dodgson, "Condensation of determinants," *Proceedings of the Royal Society*, **15**, 150–155 (1866).
4. A. C. Aitken, "On the evaluation of determinants, the formation of their adjoints, and the practical solution of simultaneous linear equations," *Proceedings of the Edinburgh Mathematical Society*, Series **2**, **3**, 207–219 (1933).

EXERCISES

1. Work exercise 1 of Chapter 4 by the method of this chapter.
2. Work exercise 2 of Chapter 4 by the method of this chapter.
3. Work exercise 4 of Chapter 4 by the method of this chapter.
4. Work exercise 10 of Chapter 4 by the method of this chapter.
5. Use the method of section 5.11 in solving the equations

$$1.569x_1 + 0.329x_2 - 0.038x_3 - 0.319x_4 = 0.511$$

$$0.555x_1 - 0.112x_2 + 0.667x_3 + 0.418x_4 = -0.629$$

$$0.372x_1 + 0.013x_2 + 1.232x_3 - 0.819x_4 = 0.713$$

$$0.842x_1 - 0.314x_2 - 0.939x_3 + 1.009x_4 = 0.338.$$

Obtain the value of x_i to four decimal places and check the results.

6. Find the values of x_i exactly for the equations

$$1.0x_1 + 0.4x_2 + 0.5x_3 + 0.6x_4 = 1.0$$

$$0.4x_1 + 1.0x_2 + 0.3x_3 + 0.4x_4 = 0.0$$

$$0.5x_1 + -0.3x_2 + 1.0x_3 + 0.2x_4 = 0.0$$

$$0.6x_1 + 0.4x_2 + 0.2x_3 + 1.0x_4 = 0.0$$

by the method of this chapter.

CHAPTER 6

The Solution of Equations with Approximate Methods

6.1 Introduction. The methods presented in Chapters 4 and 5, though they yield exact forward solutions for all problems expressible in digital numbers and can be made to yield back solutions with the use of fractional forms, have two serious drawbacks.

(*a*) Because machine capacity is soon exhausted we are practically forced to use approximations, as explained in section 4.13, if the coefficients have many digits or if there are very many equations. In this event the methods of Chapter 5 are preferable to those of Chapter 4.

(*b*) The second disadvantage of an exact method results from the fact that the compact presentation of the methods demands operations that are, with present machines, recording units rather than operational units. Actually results must be transferred regularly from the products register to the setting mechanism by hand. These manual operations take time, introduce the possibility of mistakes, and in general interfere with the smoothness and speed of the computational technique.

This second disadvantage can be overcome by using the division operation rather than multiplication as the basis of elimination. Indeed, division methods are commonly used in the direct reduction of simultaneous equations with condensation methods. Remarkable speed can be obtained with the use of an improved version of some division method such as the Doolittle method or the square root method since each recorded value is the result of an operational unit and can be computed without clearing the products register. We need not bother with numbers of many digits, since we can compute each term in the solution to the desired number of places (within machine capacity) by using incomplete numbers.

It is admitted, of course, that a division method, unless the division is exact as in the method of Chapter 5, forces us immediately into the use of approximate digital numbers. The possibility of obtaining an exact solution is lost as soon as the first (division) steps are taken. This is not so serious where the problem is itself an approximate problem.

This introduction of approximate operations is not, as at first appears, in direct violation of the recommendations of section 3.4. Here the calculational design calls for the deliberate sacrifice of the feature of exactness for the sake of ease of calculation, particularly for those problems in which exact methods eventually become impractical because of the limitations of machine capacity.

The purpose of this chapter is to present the chief condensation division methods (with variations) and with justification of each. For the present it is assumed that the coefficients are exact digital numbers. The situation where the coefficients are approximate numbers is discussed in Chapter 17.

The chief division methods outlined in this chapter are (*a*) methods of row division, (*b*) methods of diagonal division, (*c*) methods of single division, and (*d*) use of square root. Each of these, of course, has variations and can be exhibited in condensed form.

Each of these methods, too, leads us to a new function of the coefficients of the equations, as did the methods of Chapters 4 and 5. The functions of the more important methods are formally defined, and the new functions are related to those of the previous chapters so that certain important properties of these functions may be inferred from the earlier discussion. Illustrations are used throughout. The methods of Chapters 4 and 5 are suitable for work without machines (if the numbers have few digits), and indeed many of the illustrations of those chapters are selected because the detailed steps can be carried through easily without a machine. General division methods by their very nature are primarily machine methods, and no attempt is made to provide the equivalents of some of the trivial illustrations of the earlier chapters.

6.2 The methods of row division. The first method described is essentially that given in Huntington's article in the *Handbook of Mathematical Statistics* [A.1]; it might be called a method of row division. Each equation is divided by its leading coefficient, and some equation, the first, for instance, is subtracted in turn from each of the $p - 1$ others, giving a new set of $p - 1$ equations in $p - 1$ unknowns. The process is continued until one equation in one unknown is obtained. The back solution, which is performed next, does not involve divisions if the equations resulting from division by the leading terms are used, since the leading term is then unity.

If an equation has a leading term that is zero, it can be left as it is for the treatment in the next division operation.

The incomplete number (to four places) solution of the problem of Tables 4.8*a*, 4.9*a*, 4.10*a*, 5.8*a*, 5.9*a*, and 5.10 is given in Table 6.2*a*.

TABLE 6.2*a*

METHOD OF ROW DIVISION

x_1	x_2	x_3	x_4	a_{i5}	Check	Sum
1.0000	0.4000	0.5000	0.6000	0.2000	0.20001	2.7000
0.4000	1.0000	0.3000	0.4000	0.4000	0.40007	2.5000
0.5000	0.3000	1.0000	0.2000	0.6000	0.60002	2.6000
0.6000	0.4000	0.2000	1.0000	0.8000	0.79998	3.0000
1.0000	0.4000	0.5000	0.6000	0.2000	2.7000	2.7000
1.0000	2.5000	0.7500	1.0000	1.0000	6.2500	6.2500
1.0000	0.6000	2.0000	0.4000	1.2000	5.2000	5.2000
1.0000	0.6667	0.3333	1.6667	1.3333	5.0000	5.0000
	2.1000	0.2500	0.4000	0.8000	3.5500	3.5500
	0.2000	1.5000	−0.2000	1.0000	2.5000	2.5000
	0.2667	−0.1667	1.0667	1.1333	2.3000	2.3000
	1.0000	0.1190	0.1905	0.3810	1.6905	1.6905
	1.0000	7.5000	−1.0000	5.0000	12.5000	12.5000
	1.0000	−0.6250	3.9996	4.2493	8.6239	8.6239
		7.3810	−1.1905	4.6190	10.8095	10.8095
		−0.7440	3.8091	3.8683	6.9334	6.9334
		1.0000	−0.1613	0.6258	1.4645	1.4645
		1.0000	−5.1198	−5.1993	−9.3191	−9.3191
			−4.9585	−5.8251	−10.7836	−10.7836
−0.9306	0.0602	0.8153	1.1748			

This method is the least satisfactory (aside from its theoretical simplicity) of the different methods presented. The symmetry of the equations, if originally present, is lost with the first set of divisions. The method demands many divisions and $p(p + 1)$ rows. The answers are the four-decimal-place incomplete numbers of the last row. The check column shows that these answers constitute a satisfactory approximate

solution since the values of a_{i5} and a'_{i5} differ in absolute value by 0.00007 or less.

Brolyer and Chauncey [A.2, 3] have shown how this method may be simplified with the elimination of a considerable portion of the recording. Each of the even-numbered matrices in Table 6.2*a* is replaced by a single row placed directly under the equation from which it is obtained. The processes of division and subtraction are then performed in one operation. Then omission of the even numbered matrices and the use of the operational unit U_5 result in an abbreviated method of row division. For example, the first value in the third matrix of Table 6.2*a* may be obtained directly from the values of the first matrix with the computation of

$$\frac{1.0000}{0.4000} - \frac{0.4000}{1.0000} = 2.1000.$$

An illustration of this abbreviated method of row division is presented in Table 6.2*b*. Some of the entries in Table 6.2*b* differ slightly from those

TABLE 6.2*b*

ABBREVIATED METHOD OF ROW DIVISION

x_1	x_2	x_3	x_4	a_{i5}	Check	Sum
1.0000	0.4000	0.5000	0.6000	0.2000	0.19997	2.7000
0.4000	1.0000	0.3000	0.4000	0.4000	0.39997	2.5000
0.5000	0.3000	1.0000	0.2000	0.6000	0.59999	2.6000
0.6000	0.4000	0.2000	1.0000	0.8000	0.79994	3.0000
	2.1000	0.2500	0.4000	0.8000	3.5500	3.5500
	0.2000	1.5000	−0.2000	1.0000	2.5000	2.5000
	0.2667	−0.1667	1.0667	1.1333	2.3000	2.3000
		7.3810	−1.1905	4.6190	10.8095	10.8095
		−0.7441	3.8091	3.8684	6.9334	6.9334
			−4.9578	−5.8246	−10.7824	−10.7823
−0.9366	0.0601	0.8153	1.1748			

of Table 6.2*a* because the values of $a/b - c/d$ are rounded off after the U_5 operation is completed, whereas the values of a/b and c/d in Table 6.2*a* are rounded off separately.

No extensive discussion of the method of row division is presented here as the methods appearing later in this chapter are much more satisfactory if the use of division methods is indicated.

6.3 The methods of diagonal division. The methods of diagonal division are really methods of row division in which the division is made by the diagonal element of the row. This division reduces the equations, at each step of the elimination process, to a set of equations having unit diagonal terms. If we let $a_{ij}/a_{ii} = b_{ij}$, (4.1.2) reduces to

$$\begin{aligned} x_1 + b_{12}x_2 + b_{13}x_3 + b_{14}x_4 &= b_{15} \\ b_{21}x_1 + x_2 + b_{23}x_3 + b_{24}x_4 &= b_{25} \\ b_{31}x_1 + b_{32}x_2 + x_3 + b_{34}x_4 &= b_{35} \\ b_{41}x_1 + b_{42}x_2 + b_{43}x_3 + x_4 &= b_{45}. \end{aligned} \tag{1}$$

The equations of (1) are not subject to condensation by direct subtraction, since the coefficients of different variables are reduced to unity in the process, but it is not difficult to work out a condensation technique. If the first equation is multiplied by b_{21} and subtracted from the second equation, by b_{31} and subtracted from the third equation, by b_{41} and subtracted from the fourth equation, the x_1 is eliminated and we have

$$\begin{aligned} (1 - b_{21}b_{12})x_2 + (b_{23} - b_{21}b_{13})x_3 + (b_{24} - b_{21}b_{14})x_4 &= b_{25} - b_{21}b_{15} \\ (b_{32} - b_{31}b_{12})x_2 + (1 - b_{31}b_{13})x_3 + (b_{34} - b_{31}b_{14})x_4 &= b_{35} - b_{31}b_{15} \\ (b_{42} - b_{41}b_{12})x_2 + (b_{43} - b_{41}b_{13})x_3 + (1 - b_{41}b_{14})x_4 &= b_{45} - b_{41}b_{15}. \end{aligned} \tag{2}$$

If we now divide (2) by the diagonal terms, we get

$$\begin{aligned} x_2 + b_{23\cdot 1}x_3 + b_{24\cdot 1}x_4 &= b_{25\cdot 1} \\ b_{32\cdot 1}x_2 + x_3 + b_{34\cdot 1}x_4 &= b_{35\cdot 1} \\ b_{42\cdot 1}x_2 + b_{43\cdot 1}x_3 + x_4 &= b_{45\cdot 1} \end{aligned} \tag{3}$$

with

$$b_{ij\cdot 1} = \frac{b_{ij} - b_{i1}b_{1j}}{1 - b_{i1}b_{1i}}. \tag{4}$$

A similar reduction enables us to replace (3) by

$$\begin{aligned} x_3 + b_{34\cdot 12}x_4 &= b_{35\cdot 12} \\ b_{43\cdot 12}x_3 + x_4 &= b_{45\cdot 12}, \end{aligned} \tag{5}$$

where

$$b_{ij\cdot 12} = \frac{b_{ij\cdot 1} - b_{i2\cdot 1}b_{2j\cdot 1}}{1 - b_{i2\cdot 1}b_{2i\cdot 1}}. \tag{6}$$

In the same way

$$x_4 = b_{45\cdot 123} \tag{7}$$

with

$$b_{45\cdot 123} = \frac{b_{45\cdot 12} - b_{43\cdot 12}b_{35\cdot 12}}{1 - b_{43\cdot 12}b_{34\cdot 12}}. \tag{8}$$

This method may be extended to the solution of p linear equations, if the solution exists, with the continued use of the recursion formula

$$b_{ij\cdot(h)} = \frac{b_{ij\cdot(h-1)} - b_{ih\cdot(h-1)}b_{hj\cdot(h-1)}}{1 - b_{ih\cdot(h-1)}b_{hi\cdot(h-1)}}. \tag{9}$$

The forward solution leads to the single equation involving one unknown, and the general problem (4.1.1) eventually condenses to

$$x_p = b_{p,\,p+1\cdot(p-1)}. \tag{10}$$

The remaining values of x_i are obtained by back solution methods. By way of illustration, this method is now applied to the problem of Table 6.2a. It is not necessary to reduce the a's to b's for the first step, as this problem has diagonal terms.

TABLE 6.3a

METHOD OF DIAGONAL DIVISION

x_1	x_2	x_3	x_4	a_{i5}	Check	Sum
1.0000	0.4000	0.5000	0.6000	0.2000	0.20000	2.7000
0.4000	1.0000	0.3000	0.4000	0.4000	0.40006	2.5000
0.5000	0.3000	1.0000	0.2000	0.6000	0.60010	2.6000
0.6000	0.4000	0.2000	1.0000	0.8000	0.79990	3.0000
	0.8400	0.1000	0.1600	0.3200	1.4200	1.4200
	0.1000	0.7500	−0.1000	0.5000	1.2500	1.2500
	0.1600	−0.1000	0.6400	0.6800	1.3800	1.3800
	1.0000	0.1190	0.1905	0.3810	1.6905	1.6905
	0.1333	1.0000	−0.1333	0.6667	1.6667	1.6667
	0.2500	−0.1562	1.0000	1.0625	2.1563	2.1562
		0.9841	−0.1587	0.6159	1.4413	1.4414
		−0.1860	0.9524	0.9672	1.7336	1.7336
		1.0000	−0.1613	0.6259	1.4646	1.4647
		−0.1953	1.0000	1.0155	1.8202	1.8202
			0.9685	1.1377	2.1062	2.1063
−0.9366	0.0602	0.8154	1.1747			

An abbreviated form of the method of diagonal division may be designed with the use of operational units U_1 and U_9. It is not necessary to record the complete matrices that do not have unit diagonal terms, but the diagonal terms only. These can be placed, in parentheses, in the diagonal of the matrices with unit diagonal terms. Formula (9) is the basis of the calculation. We first compute the denominator and place it in the diagonal, as described above, after which the various $b_{ij\cdot(k)}$ are computed by dividing the numerator of (9) by the denominator of (9), now recorded in the diagonal. The illustration of this abbreviated method appears in Table 6.3*b*. In applying the row sum check

TABLE 6.3*b*

ABBREVIATED METHOD OF DIAGONAL DIVISION

x_1	x_2	x_3	x_4	a_{i5}	Check	Sum
1.0000	0.4000	0.5000	0.6000	0.2000	0.19996	2.7000
0.4000	1.0000	0.3000	0.4000	0.4000	0.40006	2.5000
0.5000	0.3000	1.0000	0.2000	0.6000	0.60007	2.6000
0.6000	0.4000	0.2000	1.0000	0.8000	0.79994	3.0000
	(0.8400)	0.1190	0.1905	0.3810	1.6905	1.6905
	0.1333	(0.7500)	−0.1333	0.6667	1.6667	1.6667
	0.2500	−0.1562	(0.6400)	1.0625	2.1563	2.1562
		(0.9841)	−0.1613	0.6259	1.4646	1.4646
		−0.1952	(0.9524)	1.0156	1.8204	1.8202
			(0.9685)	1.1748	2.1748	2.1746
−0.9367	0.0602	0.8154	1.1748			

and in carrying out the back solution, we should remember that the correct diagonal entries are unity. They are not the parenthetical values that are recorded for compactness and ease of computation.

For actual computation this method is inferior to one of the division methods described later. This method, like the method of row division, does not maintain symmetry. This method does have theoretical interest and may be very useful if the solutions of a large number of related equations are desired, in that every non-diagonal entry in the forward solution (excluding those of the check columns) is the value of some x_i in some set of equations obtained by deleting variables and equa-

tions from (4.1.1). The formal proof of this is indicated by (1), where each b_i is identified as an x_i. Thus in the illustration above, $p = 4$, so that

$$x_{45\cdot 123} = b_{45\cdot 123} = 1.1748.$$

Thus $b_{45\cdot 123}$ is a more explicit form for x_4. It is that particular x_4 that is obtained when the other variables are x_1, x_2, and x_3 and when the right-hand terms are indicated by a_{i5}.

If the variables x_3 and the next-to-last equation are omitted from (4.1.2), we have

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{14}x_4 &= a_{15} \\ a_{21}x_1 + a_{22}x_2 + a_{24}x_4 &= a_{25} \\ a_{41}x_1 + a_{42}x_2 + a_{44}x_4 &= a_{45} \end{aligned} \tag{11}$$

so that

$$x_4 = b_{45\cdot 12}.$$

Similarly, if the equations are

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= a_{14} \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= a_{24} \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= a_{34}, \end{aligned} \tag{12}$$

then

$$x_3 = b_{34\cdot 12}.$$

If the a_{ij} in (11) and (12) are identical with the a_{ij} of Table 6.3*b*, the values of $b_{45\cdot 12}$ and $b_{34\cdot 12}$, as well as others, are available by inspection from that table. Thus $b_{45\cdot 12} = 1.0156$ and $b_{34\cdot 12} = -0.1613$. In general the value of $b_{ij\cdot kl\cdots}$ is the value of x_i when the constant column is indicated by a_{*j} and the other variables involved in the equations are x_k, x_l, $\cdots$. A formal proof of this could be made by applying (10), after rearranging the variables and equations to conform to the desired order of elimination.

We can add the subscripts to x, as well as to b, to indicate the solution. Thus $x_{ij\cdot kl\cdots}$ has the same meaning as $b_{ij\cdot kl\cdots}$.

All the non-diagonal values of the abbreviated method of diagonal division are themselves parts of solutions of equations that are related to the original equations. Thus the formal solution of Table 6.3*b* exhibits the values of x_i shown in Table 6.3*c*.

It is true that the values of Table 6.3*c* do not constitute all the x_i solutions of the related equations, but the back solution method can be used to give complete sets of solutions as indicated in Chapter 8.

Formulas (4), (6), and (8), and, in general, (9) may be used as recursion formulas for obtaining the solution of equations involving h additional variables from those involving $h - 1$ additional variables. A

TABLE 6.3*c*

VALUE OF x_i

x_1	x_2	x_3	x_4	
( )	x_{12}	x_{13}	x_{14}	x_{15}
x_{21}	( )	x_{23}	x_{24}	x_{25}
x_{31}	x_{32}	( )	x_{34}	x_{35}
x_{41}	x_{42}	x_{43}	( )	x_{45}
	( )	$x_{23\cdot1}$	$x_{24\cdot1}$	$x_{25\cdot1}$
	$x_{32\cdot1}$	( )	$x_{34\cdot1}$	$x_{35\cdot1}$
	$x_{42\cdot1}$	$x_{43\cdot1}$		$x_{45\cdot1}$
		( )	$x_{34\cdot12}$	$x_{35\cdot12}$
		$x_{43\cdot12}$	( )	$x_{45\cdot12}$
			( )	$x_{45\cdot123}$

classical formula [B, p. 142] is the equivalent of (9). The notation of (9) is not identical with that of the classical formula since this uses the second primary subscript to indicate the variable and the first primary subscript to indicate the right side of the equation. The notation used in (9) is consistent with that usually followed by mathematicians in that the first subscript indicates the row of the matrix and the second subscript indicates the column.

The classical recursion method is essentially the same as this method of diagonal division, and the computational form outlined above may be used to replace the repeated application of the recursion formula.

The method cannot be continued, at least not without modification, if any diagonal term is zero. One modification is to shift to the method of multiplication and subtraction and follow through as in Chapter 4. With this modification the general rules with reference to equivalent, consistent, and homogeneous equations are identical with those of Chapter 4. Zero values may be approximate. We must be careful

that we do not record a value as specific when it is nothing but 0/0 in the form $(0 + \epsilon_1)/(0 + \epsilon_2)$. A similar remark holds for $(k + \epsilon_1)/(0 + \epsilon_2)$. The equations are inconsistent if $k \neq 0$. There is some difficulty in absolute identification of these cases; this is the price we must pay for approximate (division) methods.

In relating the b's to the m's, we note that in so far as the theory is concerned we may assume that the operations are exact and that no approximations enter. Thus (9) is exact. The approximation arises only when we apply it to a numerical problem. Now

$$b_{ij} = \frac{a_{ij}}{a_{ii}}.$$

Similarly,

$$b_{ij\cdot 1} = \frac{b_{ij} - b_{i1}b_{1j}}{1 - b_{i1}b_{1i}} = \frac{\dfrac{a_{ij}}{a_{ii}} - \dfrac{a_{i1}}{a_{ii}}\dfrac{a_{1j}}{a_{11}}}{1 - \dfrac{a_{i1}}{a_{ii}}\dfrac{a_{1i}}{a_{11}}} = \frac{m_{ij\cdot 1}}{m_{ii\cdot 1}} = \frac{d_{ij\cdot 1}}{d_{ii\cdot 1}}$$

$$b_{ij\cdot 12} = \frac{m_{ij\cdot 12}}{m_{ii\cdot 12}} = \frac{d_{ij\cdot 12}}{d_{ii\cdot 12}}$$

$$b_{ij\cdot 123} = \frac{m_{ij\cdot 123}}{m_{ii\cdot 123}} = \frac{d_{ij\cdot 123}}{d_{ii\cdot 123}}$$

and, in general,

$$b_{ij\cdot(h)} = \frac{m_{ij\cdot(h)}}{m_{ii\cdot(h)}} = \frac{d_{ij\cdot(h)}}{d_{ii\cdot(h)}}. \tag{13}$$

This might be established less formally since each of these elimination processes leads to the same value $x_{ij\cdot(h)}$.

6.4 The methods of single division. The methods of this section are characterized by division by the leading element for one of the p equations. The equation resulting from this division is then combined with each of the other $p - 1$ equations in turn to give $p - 1$ new equations in $p - 1$ unknowns. It is conventional to use the first equation as the one to be divided by its leading coefficient. The new equation, which in effect replaces the equation that was divided, may then be placed at the bottom of the p equations. There is no great loss of generality, and a somewhat simpler technique results if the variables are eliminated in 1, 2, 3, 4 order, although a slightly more involved computational technique can be worked out, as in Chapter 4, when some other order,

or a non-diagonal pivot, is used. The coefficient serving as a pivot should be different from zero.

By dividing the first equation by a_{11} and letting $a_{1i}/a_{11} = b_{1i}$ as in (6.3.1), the equations (4.1.2) become

$$\begin{aligned} a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + a_{24}x_4 &= a_{25} \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + a_{34}x_4 &= a_{35} \\ a_{41}x_1 + a_{42}x_2 + a_{43}x_3 + a_{44}x_4 &= a_{45} \\ x_1 + b_{12}x_2 + b_{13}x_3 + b_{14}x_4 &= b_{15}. \end{aligned} \tag{1}$$

Multiply the last equation by a_{21} and subtract from the first equation, by a_{31} and subtract from the second equation, by a_{41} and subtract from the third equation, and get the three equations

$$\begin{aligned} g_{22\cdot1}x_2 + g_{23\cdot1}x_3 + g_{24\cdot1}x_4 &= g_{25\cdot1} \\ g_{32\cdot1}x_2 + g_{33\cdot1}x_3 + g_{34\cdot1}x_4 &= g_{35\cdot1} \\ g_{42\cdot1}x_2 + g_{43\cdot1}x_3 + g_{44\cdot1}x_4 &= g_{45\cdot1} \end{aligned} \tag{2}$$

with

$$g_{ij\cdot1} = a_{ij} - a_{i1}b_{1j}. \tag{3}$$

Now

$$\frac{g_{ij\cdot1}}{g_{ii\cdot1}} = \frac{a_{ij} - a_{i1}b_{1j}}{a_{ii} - a_{i1}b_{1i}} = \frac{\dfrac{a_{ij}}{a_{ii}} - \dfrac{a_{i1}}{a_{ii}}b_{1j}}{\dfrac{a_{ii}}{a_{ii}} - \dfrac{a_{i1}}{a_{ii}}b_{1i}} = \frac{b_{ij} - b_{i1}b_{1j}}{1 - b_{i1}b_{1i}} = b_{ij\cdot1} \tag{4}$$

as defined by (6.3.4). We divide the first equation of (2) by $g_{22\cdot1}$ and place the results in the bottom row to get

$$\begin{aligned} g_{32\cdot1}x_2 + g_{33\cdot1}x_3 + g_{34\cdot1}x_4 &= g_{35\cdot1} \\ g_{42\cdot1}x_2 + g_{43\cdot1}x_3 + g_{44\cdot1}x_4 &= g_{45\cdot1} \\ x_2 + b_{23\cdot1}x_3 + b_{24\cdot1}x_4 &= b_{25\cdot1}. \end{aligned} \tag{5}$$

We eliminate as before and obtain

$$\begin{aligned} g_{33\cdot12}x_3 + g_{34\cdot12}x_4 &= g_{35\cdot12} \\ g_{43\cdot12}x_3 + g_{44\cdot12}x_4 &= g_{45\cdot12} \end{aligned} \tag{6}$$

with

$$g_{ij\cdot 12} = g_{ij\cdot 1} - g_{i2\cdot 1}b_{2j\cdot 1}. \tag{7}$$

We note that

$$\frac{g_{ij\cdot 12}}{g_{ii\cdot 12}} = b_{ij\cdot 12} \tag{8}$$

and get, after dividing the first equation of (6) by $g_{33\cdot 12}$,

$$\begin{aligned} g_{43\cdot 12}x_3 + g_{44\cdot 12}x_4 &= g_{45\cdot 12} \\ x_3 + b_{34\cdot 12}x_4 &= b_{35\cdot 12}. \end{aligned} \tag{9}$$

The next step yields

$$g_{44\cdot 123}x_4 = g_{45\cdot 123} \tag{10}$$

with

$$g_{ij\cdot 123} = g_{ij\cdot 12} - g_{i3\cdot 12}b_{3j\cdot 12} \tag{11}$$

and, finally,

$$x_4 = x_{45\cdot 123} = \frac{g_{45\cdot 123}}{g_{44\cdot 123}} = b_{45\cdot 123}. \tag{12}$$

The back solution is readily accomplished by substituting in the equations having the b's as coefficients.

A symmetric form of the solution is indicated in Table 6.4a. Because of the importance of this method the general notation and the numerical illustration are presented. A forward solution is carried out with a series of U_1 operations, followed by a series of divisions. The back solution is carried out with a series of U_3 operations.

Effective abbreviations for the method of single division are available. We record only the elements of the first row and the first column at each step since

$$\begin{aligned} g_{ij\cdot 1} &= a_{ij} - a_{i1}b_{1j} \\ g_{ij\cdot 12} &= g_{ij\cdot 1} - g_{i2\cdot 1}b_{2j\cdot 1} \\ &= a_{ij} - g_{i1}b_{1j} - g_{i2\cdot 1}b_{2j\cdot 1} \\ g_{ij\cdot 123} &= g_{ij\cdot 12} - g_{i3\cdot 12}b_{3j\cdot 12} \\ &= a_{ij} - a_{i1}b_{1j} - g_{i2\cdot 1}b_{2j\cdot 1} - g_{i3\cdot 12}b_{3j\cdot 12} \\ &\quad \text{etc.} \end{aligned} \tag{13}$$

TABLE 6.4*a*

METHOD OF SINGLE DIVISION—SYMMETRIC

General

x_1	x_2	x_3	x_4	a_{i5}	Check	Sum
a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a'_{15}	a_{1T}
*	a_{22}	a_{23}	a_{24}	a_{25}	a'_{25}	a_{2T}
*	*	a_{33}	a_{34}	a_{35}	a'_{35}	a_{3T}
*	*	*	a_{44}	a_{45}	a'_{45}	a_{4T}
1	b_{12}	b_{13}	b_{14}	b_{15}	sum	b_{1T}
	$g_{22\cdot1}$	$g_{23\cdot1}$	$g_{24\cdot1}$	$g_{25\cdot1}$	sum	$g_{2T\cdot1}$
	*	$g_{33\cdot1}$	$g_{34\cdot1}$	$g_{35\cdot1}$	sum	$g_{3T\cdot1}$
	*	*	$g_{44\cdot1}$	$g_{45\cdot1}$	sum	$g_{4T\cdot1}$
	1	$b_{23\cdot1}$	$b_{24\cdot1}$	$b_{25\cdot1}$	sum	$b_{2T\cdot1}$
		$g_{33\cdot12}$	$g_{34\cdot12}$	$g_{35\cdot12}$	sum	$g_{3T\cdot12}$
		*	$g_{44\cdot12}$	$g_{45\cdot12}$	sum	$g_{4T\cdot12}$
		1	$b_{34\cdot12}$	$b_{35\cdot12}$	sum	$b_{3T\cdot12}$
			$g_{44\cdot123}$	$g_{45\cdot123}$	sum	$g_{4T\cdot123}$
			1	$b_{45\cdot123}$	sum	$b_{4T\cdot123}$
$b_{15\cdot234}$	$b_{25\cdot134}$	$b_{35\cdot124}$	$b_{45\cdot123}$			
$b_{1T\cdot234}$	$b_{2T\cdot134}$	$b_{3T\cdot124}$	$b_{4T\cdot123}$			

Illustration

x_1	x_2	x_3	x_4	a_{i5}	Check	Sum
1.0000	0.4000	0.5000	0.6000	0.2000	0.19996	2.7000
*	1.0000	0.3000	0.4000	0.4000	0.40004	2.5000
*	*	1.0000	0.2000	0.6000	0.59992	2.6000
*	*	*	1.0000	0.8000	0.79996	3.0000
1.0000	0.4000	0.5000	0.6000	0.2000	2.70000	2.7000
	0.8400	0.1000	0.1600	0.3200	1.4200	1.4200
	*	0.7500	−0.1000	0.5000	1.2500	1.2500
	*	*	0.6400	0.6800	1.3800	1.3800
	1.0000	0.1190	0.1905	0.3810	1.6905	1.6905
		0.7381	−0.1190	0.4619	1.0810	1.0810
		*	0.6095	0.6190	1.1095	1.1095
		1.0000	−0.1612	0.6258	1.4646	1.4646
			0.5903	0.6935	1.2838	1.2838
			1.0000	1.1748	2.1748	2.1748
−0.9366	0.0602	0.8152	1.1748			
0.0634	1.0602	1.8152	2.1748			

The values of $g_{ij\cdot(h)}$ may be computed as U_3 operational units after the b's have been computed.

The important calculational formulas (13) are essentially due to Gauss [C.1], who used a different notation and applied them to symmetric (least squares) problems. They are also the basis of the Doolittle method [C.2]. The notation here used is similar to that used by Yule in presenting the theory of multiple and partial correlation. It seems preferable to Gauss's theory since the results can be immediately translated to correlation and regression theory.

Like the method of multiplication and subtraction, the method of single division has been developed independently by many authors. It

TABLE 6.4*b*

ABBREVIATED METHOD OF SINGLE DIVISION

x_1	x_2	x_3	x_4	$a_{.5}$	Sum	Check
1.0000	0.4000	0.5000	0.6000	0.2000	0.19996	2.7000
0.4000	1.0000	0.3000	0.4000	0.4000	0.40004	2.5000
0.5000	0.3000	1.0000	0.2000	0.6000	0.59992	2.6000
0.6000	0.4000	0.2000	1.0000	0.8000	0.79996	3.0000
1.0000	0.4000	0.5000	0.6000	0.2000	2.70000	2.7000
	0.8400	0.1000	0.1600	0.3200	1.4200	1.4200
	0.1000	*	*	*	*	*
	0.1600	*	*	*	*	*
	1.0000	0.1190	0.1905	0.3810	1.6905	1.6905
		0.7381	−0.1190	0.4619	1.0810	1.0810
		−0.1190	*	*	*	*
		1.0000	−0.1612	0.6258	1.4646	1.4646
			0.5903	0.6935	1.2838	1.2838
			1.0000	1.1748	2.1748	2.1748
−0.9366	0.0602	0.8152	1.1748			

appears that Gauss and Doolittle applied the method only to symmetric equations. More recent authors, for example, Aitken, Banachiewicz, Dwyer, and Crout [C.3–8], have emphasized the use of the method, or variations of it, in connection with non-symmetric problems.

An abbreviated form of the method is illustrated in Table 6.4*b*, where the computational technique is based on (13). A still more condensed

form is based on the fact that

$$b_{ij\cdot 1} = \frac{a_{ij} - a_{i1}b_{1j}}{a_{11}}$$

$$b_{ij\cdot 12} = \frac{a_{ij} - a_{i1}b_{1j} - g_{i2\cdot 1}b_{2j\cdot 1}}{g_{22\cdot 1}} \tag{14}$$

$$b_{ij\cdot 123} = \frac{a_{ij} - a_{i1}b_{1j} - g_{i2\cdot 1}b_{2j\cdot 1} - g_{i3\cdot 12}b_{3j\cdot 12}}{g_{33\cdot 12}}$$

etc.

are U_9 operations. It is hence necessary only to fill in the first row and the first column of each matrix. The computational form is illustrated in Table 6.4c. Slight variations result if the numerator is rounded before division.

TABLE 6.4c

CONDENSED ABBREVIATED METHOD OF SINGLE DIVISION

x_1	x_2	x_3	x_4	a_{i5}	Check	Sum
1.0000	0.4000	0.5000	0.6000	0.2000	0.19996	2.7000
0.4000	1.0000	0.3000	0.4000	0.4000	0.40004	2.5000
0.5000	0.3000	1.0000	0.2000	0.6000	0.59992	2.6000
0.6000	0.4000	0.2000	1.0000	0.8000	0.79996	3.0000
1.0000	0.4000	0.5000	0.6000	0.2000	2.7000	2.7000
	(0.8400)	0.1190	0.1905	0.3810	1.6905	1.6905
	0.1000	*	*	*	*	*
	0.1600	*	*	*	*	*
		(0.7381)	−0.1612	0.6258	1.4646	1.4646
		−0.1190	*	*	*	*
			(0.5903)	1.1748	2.1748	2.1748
−0.9366	0.0602	0.8152	1.1748			

A useful form of the abbreviated method of single division results from a rearrangement of the entries of Table 6.4d. This is the computational form advocated in Poland in 1938 by Banachiewicz [C.7] and in 1941 in this country by P. D. Crout [C.8]. The last row of the forward solution of Table 6.4c is moved into the vacant spaces in the row above it. The entries of this new row and those of the row above it are moved

into the vacant two rows of the matrix above, etc. The last row of the first matrix and the values of a_{i1} are placed in the first column of the second matrix. The general construction of the computing form appears in Table 6.4*d*. This computing form is so useful that it is described in some detail.

TABLE 6.4*d*

COMPACT FORM OF THE METHOD OF SINGLE DIVISION

x_1	x_2	x_3	x_4	a_{i5}	Check	Sum
a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a'_{15}	a_{1T}
a_{21}	a_{22}	a_{23}	a_{24}	a_{25}	a'_{25}	a_{2T}
a_{31}	a_{32}	a_{33}	a_{34}	a_{35}	a'_{35}	a_{3T}
a_{41}	a_{42}	a_{43}	a_{44}	a_{45}	a'_{45}	a_{4T}
1						
a_{11}	b_{12}	b_{13}	b_{14}	b_{15}	sum	b_{1T}
	1					
a_{21}	$g_{22\cdot1}$	$b_{23\cdot1}$	$b_{24\cdot1}$	$b_{25\cdot1}$	sum	$b_{2T\cdot1}$
		1				
a_{31}	$g_{32\cdot1}$	$g_{33\cdot12}$	$b_{34\cdot12}$	$b_{35\cdot12}$	sum	$b_{3T\cdot12}$
			1			
a_{41}	$g_{42\cdot1}$	$g_{43\cdot12}$	$g_{44\cdot123}$	$b_{45\cdot123}$	sum	$b_{4T\cdot123}$
$x_{15\cdot234}$	$x_{25\cdot134}$	$x_{35\cdot124}$	$x_{45\cdot123}$			

(*a*) Write the equations in the usual synthetic form and provide a check column.
(*b*) Draw a line through the diagonal of the second matrix.
(*c*) Divide the first row of the first matrix by its leading term to obtain the first row of the second matrix.
(*d*) Write the first column of the first matrix as the first column of the second matrix.
(*e*) Calculate the values of $g_{i2\cdot1}$ and record them in the second column of the second matrix.

This computational procedure is simple in that it requires only the mental transfer of the value of a_{i2} to the corresponding position in the second matrix and the subtraction of the product of the term at the left (a_{i1}) and the term at the top (b_{12}).

(*f*) Calculate the values $b_{2j\cdot1} = g_{2j\cdot1}/g_{22\cdot1}$ and record them in the second row of the second matrix. These calculations are similar

to the $g_{i2\cdot1}$ operations, with an additional division by $g_{22\cdot1}$. The operation is a U_9 operational unit.

(*g*) Calculate the values $g_{i3\cdot12} = a_{i3} - a_{i1}b_{13} - g_{i2\cdot1}b_{23\cdot1}$ and enter them in the third column of the second matrix. The technique of calculation is an extension of that of (*f*), which features the products of terms in the *i*th row and third column.

(*h*) Calculate the values $b_{3j\cdot12} = (a_{3j} - a_{31}b_{1j} - g_{32\cdot1}b_{2j\cdot1})/g_{22\cdot1}$ and record in the third row of the second matrix.

(*i*) Continue the process until the forward solution is complete.

(*j*) Carry out the back solution with the use of the *b*'s and the unit diagonal terms.

(*k*) Substitute the proposed solution in each of the original equations (which is easily accomplished in the symmetric form, using U_4 operational units), write the answers in the check column, and see how well these incomplete numbers satisfy the original equations.

The method is applied to the four-variable problem previously used, in Table 6.4*e*. The unit diagonal terms are omitted in the printed version.

TABLE 6.4*e*

ILLUSTRATION

x_1	x_2	x_3	x_4	a_{i5}	Check	Sum
1.0000	0.4000	0.5000	0.6000	0.2000	0.19996	2.7000
0.4000	1.0000	0.3000	0.4000	0.4000	0.40004	2.5000
0.5000	0.3000	1.0000	0.2000	0.6000	0.59992	2.6000
0.6000	0.4000	0.2000	1.0000	0.8000	0.79996	3.0000
(1.0000)	0.4000	0.5000	0.6000	0.2000	2.7000	2.7000
0.4000	(0.8400)	0.1190	0.1905	0.3810	1.6905	1.6905
0.5000	0.1000	(0.7381)	−0.1612	0.6258	1.4646	1.4646
0.6000	0.1600	−0.1190	(0.5903)	1.1748	2.1748	2.1748
−0.9366	0.602	0.8152	1.1748			

An alternative method calls for the use of *a*'s and *g*'s in the row and *b*'s in the column. This has been used in a recent paper [C]. A check for column sums, as well as row sums, may be added if desired.

The computing form in Tables 6.4*d* and 6.4*e* is very similar to the computing form of Tables 4.9*b* and 5.9*b*. All these forms can be reduced to

one basic form through the use of formulas relating the b's, g's, m's, and d's.

This method preserves symmetry if originally present. Thus

$$(15) \qquad g_{ij\cdot 1} = a_{ij} - a_{i1}b_{1j} = a_{ij} - \frac{a_{i1}a_{1j}}{a_{11}} = a_{ji} - \frac{a_{j1}a_{1i}}{a_{11}} = g_{ji\cdot 1}$$

if $a_{ij} = a_{ji}$, and this can be extended to the general case.

This property of symmetry can be used to lighten the calculation in the compact form of the method of single division since the values of $b_{ij\cdot(h)}$ can be computed at once from the values of $g_{ji\cdot(h)}$ with the use of the formula

$$(16) \qquad b_{ij\cdot(h)} = \frac{g_{ij\cdot(h)}}{g_{ii\cdot(h)}} = \frac{g_{ji\cdot(h)}}{g_{ii\cdot(h)}}.$$

The formulas (16) can be written

$$(17) \qquad \begin{aligned} g_{ij\cdot 1} &= \begin{cases} a_{ij} - a_{1i}b_{1j} \\ a_{ij} - a_{1j}b_{1i} \end{cases} \\ g_{ij\cdot 12} &= \begin{cases} a_{ij} - a_{1i}b_{1j} - g_{2i\cdot 1}b_{2j\cdot 1} \\ a_{ij} - a_{1j}b_{1i} - g_{2j\cdot 1}b_{2i\cdot 1} \end{cases} \\ g_{ij\cdot 123} &= \begin{cases} a_{ij} - a_{1i}b_{1j} - g_{2i\cdot 1}b_{2j\cdot 1} - g_{3i\cdot 12}b_{3j\cdot 12} \\ a_{ij} - a_{1j}b_{1i} - g_{2j\cdot 1}b_{2i\cdot 1} - g_{3j\cdot 12}b_{3i\cdot 12}, \end{cases} \end{aligned}$$

etc.

These formulas can be applied to the method of single division. None of the entries below the diagonal need be recorded since each is equal to an entry above the diagonal. The elements to be multiplied are at the top and the bottom of the ith and jth columns, respectively. An illustration is given in the earlier Table 6.4a.

Abbreviated single division methods can also be applied to symmetric problems. For this purpose the equations (17) are used with the computation of a single row of g's and a single row of b's at each elimination. Some slight modification makes for ease of computation. The first row of b's is detached from the first set of equations and placed under the first equation (which is written the second time) to form the first equation doublet. The values of $g_{2j\cdot 1}$ are then computed with the use of $g_{2j\cdot 1} = a_{2j} - a_{1j}b_{12}$ (or $g_{2j\cdot 1} = a_{2j} - a_{j2}b_{1j}$). These values are divided by $g_{22\cdot 1}$ to get $b_{2j\cdot 1}$. This process is continued through the forward solution, and the back solution is carried out in the usual manner.

The resultant solution is essentially an abbreviated and slightly modified form of the Doolittle solution, which has been advocated for several decades as a good method for solving symmetric simultaneous equations. Doolittle's solution was arrived at by using multiplication tables and hence had many entries that are not necessary for modern computing machines, although Doolittle understood that these entries were auxiliary to the main solution [C.2]. It also made use of division by the negative of the diagonal term. In earlier work it was considered necessary to present the advantages of this abbreviation and modification of the conventional Doolittle method, but these seem now to be generally accepted.

This abbreviated Doolittle method, which may be considered as an abbreviated form of the method of single division applied to symmetric problems, is presented in Table 6.4*f*. Comparison of the results of

TABLE 6.4*f*

ABBREVIATED DOOLITTLE METHOD

x_1	x_2	x_3	x_4	a_{i5}	Check	Sum
1.0000	0.4000	0.5000	0.6000	0.2000	0.19996	2.7000
*	1.0000	0.3000	0.4000	0.4000	0.40004	2.5000
*	*	1.0000	0.2000	0.6000	0.59992	2.6000
*	*	*	1.0000	0.8000	0.79996	3.0000
1.0000	0.4000	0.5000	0.6000	0.2000	2.7000	2.7000
1.0000	0.4000	0.5000	0.6000	0.2000	2.7000	2.7000
	0.8400	0.1000	0.1600	0.3200	1.4200	1.4200
	1.0000	0.1190	0.1905	0.3810	1.6905	1.6905
		0.7381	−0.1190	0.4619	1.0810	1.0810
		1.0000	−0.1612	0.6258	1.4646	1.4646
			0.5903	0.6935	1.2838	1.2838
			1.0000	1.1748	2.1748	2.1748
−0.9366	0.0602	0.8152	1.1748			
0.0634	1.0602	1.8152	2.1748			

Table 6.4*f* with the results of Table 6.4*d* reveals that the computations (for symmetric matrices) are indeed identical. Each b in the abbreviated Doolittle method may be computed along with the g as a single operational unit with dual recording just as in Table 6.4*d*, when the matrix is symmetric. The earlier presentation is more compact and is

easy to operate when the number of equations is small, but the row-by-column multiplication becomes awkward with large matrices. The Doolittle solution is probably preferred in that we can accomplish the multiplication with less effort and fewer mistakes if the numbers to be multiplied can be identified more easily.

Waugh published an abbreviated form of the Doolittle method in 1935 [C.3]. Banachiewicz published the method of decomposition [C.7] that is applicable to non-symmetric matrices as well as symmetric matrices in Poland in 1938. Dwyer related the abbreviated Doolittle method to other methods of solving equations in 1941 [C.4], and Crout [C.8] presented his solution, applicable to non-symmetric problems, in 1941. The Crout solution is similar to one of the Banachiewicz solutions. Banachiewicz, over a decade ago, not only provided good techniques, but he also saw the point stressed later by other writers [D], that the basic problem is really one of matrix factorization, or "decomposition" as he called it.

As indicated above, the column-by-column multiplication of the Gauss-Doolittle solution seems to be preferred to the row-by-column multiplication of the compact method of single division for many variables, and yet the Gauss-Doolittle solution is applicable to symmetric problems only. It is possible to transpose columns with rows so as to carry out the calculation with a column-by-column multiplication even when the matrix is not symmetric. A description of the solution is given below. See also [E].

(*a*) Present the problem in the first p rows.

(*b*) Write the transpose of the first column of the first matrix as the first row of the first doublet. Form the row sum check.

(*c*) Divide the first row by its leading coefficient and write the resulting b_{1j} in the second row of the first doublet. Use a row sum check.

(*d*) Compute the successive values $g_{i2\cdot 1}$ by the formula $g_{i2\cdot 1} = a_{i2} - a_{i1}b_{12}$ and record them in the ith column of the first row of the second doublet. Each of these values is obtained by taking the a_{i2} in the first matrix and subtracting from it the product of the b element directly below it and the a element in the ith column and first row of the first doublet. This can be done by taking the b directly under the element and by taking the a under the diagonal term of the same row.

(*e*) Compute the values $b_{2j\cdot 1} = (a_{2j} - a_{21}b_{1j})/g_{22\cdot 1}$ and record each in the jth column of the second row of the doublet. They are obtained by subtracting from a_{2j} the product of the b directly under the entry with the a term and dividing by $g_{22\cdot 1}$.

(*f*) This general process is continued.

The outline of the solution is shown in Table 6.4*g* and the application to the non-symmetric illustration of Table 4.13*a* in Table 6.4*h*. The solution using the compact method of single division is placed in Table 6.4*i* for comparison. Note that the computations are the same, but that the arrangements are different.

TABLE 6.4*g*

GENERALIZED GAUSS-DOOLITTLE SOLUTION

x_1	x_2	x_3	x_4		Check	Sum
a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a'_{15}	a_{1T}
a_{21}	a_{22}	a_{23}	a_{24}	a_{25}	a'_{25}	a_{2T}
a_{31}	a_{32}	a_{33}	a_{34}	a_{35}	a'_{35}	a_{3T}
a_{41}	a_{42}	a_{43}	a_{44}	a_{45}	a'_{45}	a_{4T}
a_{T1}	a_{T2}	a_{T3}	a_{T4}			
a_{11}	a_{21}	a_{31}	a_{41}		sum	a_{T1}
1	b_{12}	b_{13}	b_{14}	b_{15}	sum	b_{1T}
	$g_{22\cdot1}$	$g_{32\cdot1}$	$g_{42\cdot1}$		sum	$g_{T2\cdot1}$
	1	$b_{23\cdot1}$	$b_{24\cdot1}$	$b_{25\cdot1}$	sum	$b_{2T\cdot1}$
		$g_{33\cdot12}$	$g_{43\cdot12}$		sum	$g_{T3\cdot12}$
		1	$b_{34\cdot12}$	$b_{35\cdot12}$	sum	$b_{3T\cdot12}$
			$g_{44\cdot123}$		sum	$g_{T4\cdot123}$
			1	$b_{45\cdot123}$	sum	$b_{4T\cdot123}$
$b_{15\cdot234}$	$b_{25\cdot134}$	$b_{35\cdot124}$	$b_{45\cdot123}$			

In a Doolittle solution with many variables we may cover the eliminated rows and columns of the original matrix with rulers, blotters, a cardboard, or template with a right-angle cut.

The basic method of single division, of which these solutions are all special cases, is such a simple method that it appears to have been developed independently by many authors. Aitken [C.6], who has done much to popularize its use, has called it the method of pivotal condensation; others have called it the method of elimination. The author prefers to call it the method of single division in that only one row of each matrix is divided. The title "method of pivotal condensation" does not seem to distinguish it sufficiently from other direct pivotal methods (both exact and approximate) that are discussed in this chapter and in the two

previous chapters. A similar remark holds for "the method of elimination."

Some of the latest fully automatic machines have been equipped with a device that enables us to disconnect the revolutions register while products are being computed. Machines equipped with this device,

TABLE 6.4*h*

GENERALIZED GAUSS-DOOLITTLE SOLUTION
WAUGH-DWYER ILLUSTRATION

x_1	x_2	x_3	x_4	a_{i5}	Check	Sum
26	−10	15	32	23	23.00042	86
19	45	−14	−8	57	56.99947	99
−12	16	27	13	47	46.99997	91
32	29	−35	28	−68	−68.00001	−14
65	80	−7	65			
26.00000	19.00000	−12.00000	32.00000		65.0000	65.00000
1.00000	−0.38462	0.57692	1.23077	0.88462	3.30769	3.30769
	52.30778	11.38456	41.30784		105.00018	105.00030
	1.00000	−0.47720	−0.60000	0.76838	0.69118	0.69118
		39.35575	−33.74934		5.60651	5.60634
		1.00000	0.87916	1.24169	3.12085	3.12085
			43.07113		43.07113	43.07113
			1.00000	−1.99999	−0.99999	−0.99999
2.00000	0.99999	3.00000	−1.99999			

TABLE 6.4*i*

COMPACT METHOD OF SINGLE DIVISION

x_1	x_2	x_3	x_4	a_{i5}	Check	Sum
26	−10	15	32	23	23.00042	86
19	45	−14	−8	57	56.99947	99
−12	16	27	13	47	46.99997	91
32	29	−35	28	−68	−68.00001	−14
26	−0.38462	0.57692	1.23077	0.88462	3.30769	3.30769
19	52.30778	−0.47720	−0.60000	0.76838	0.69118	0.69118
−12	11.38456	39.35575	0.87916	1.24169	3.12085	3.12085
32	41.30784	−33.74934	43.07113	−1.99999	−0.99999	−0.99999
2.00000	0.99999	3.00000	−1.99999			

as well as the usual device for positive and negative cumulation of a quotient, are able to compute the operational unit

$$U_{14} = \frac{ab}{c} \pm \frac{de}{f} \pm \frac{hi}{j} \pm, \text{ etc.}$$

A further modification of the method of single division for the symmetric case is now available; it seems to be new. Use (17) with the b's

TABLE 6.4j

COMPACT FORM OF THE METHOD OF SINGLE DIVISION—SYMMETRIC

x_1	x_2	x_3	x_4	a_{15}	Check	Sum
a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a'_{15}	a_{1T}
*	a_{22}	a_{23}	a_{24}	a_{25}	a'_{25}	a_{2T}
*	*	a_{33}	a_{34}	a_{35}	a'_{35}	a_{3T}
*	*	*	a_{44}	a_{45}	a'_{45}	a_{4T}
a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	sum	a_{1T}
	$g_{22\cdot1}$	$g_{23\cdot1}$	$g_{24\cdot1}$	$g_{25\cdot1}$	sum	$g_{2T\cdot1}$
		$g_{33\cdot12}$	$g_{34\cdot12}$	$g_{35\cdot12}$	sum	$g_{3T\cdot12}$
			$g_{44\cdot123}$	$g_{45\cdot123}$	sum	$g_{4T\cdot123}$
$x_{15\cdot234}$	$x_{25\cdot134}$	$x_{35\cdot124}$	$x_{45\cdot123}$			

ILLUSTRATION

x_1	x_2	x_3	x_4	a_{15}	Check	Sum
1.0000	0.4000	0.5000	0.6000	0.2000	0.19996	2.7000
*	1.0000	0.3000	0.4000	0.4000	0.40004	2.5000
*	*	1.0000	0.2000	0.6000	0.59992	2.6000
*	*	*	1.0000	0.8000	0.79996	3.0000
1.0000	0.4000	0.5000	0.6000	0.2000	2.7000	2.7000
	0.8400	0.1000	0.1600	0.3200	1.4200	1.4200
		0.7381	−0.1190	0.4619	1.0810	1.0810
			0.5903	0.6935	1.2838	1.2838
−0.9336	0.0602	0.8152	1.1748			

replaced by their expansions in terms of the a's and g's to get

$$g_{ij\cdot 1} = a_{ij} - \frac{a_{1i}a_{1j}}{a_{11}}$$

$$g_{ij\cdot 12} = a_{ij} - \frac{a_{1i}a_{1j}}{a_{11}} - \frac{g_{2i\cdot 1}g_{2j\cdot 1}}{g_{22\cdot 1}} \tag{18}$$

$$g_{ij\cdot 123} = a_{ij} - \frac{a_{1i}a_{1j}}{a_{11}} - \frac{g_{2i\cdot 1}g_{2j\cdot 1}}{g_{22\cdot 1}} - \frac{g_{3i\cdot 12}g_{3j\cdot 12}}{g_{33\cdot 12}},$$

etc.

The g's can be computed as U_{14} operational units. The b's can be eliminated from the forward solution that now exhibits the computed g's. The computational form is indicated, and application is made to the illustration of Table 6.4*f* in Table 6.4*j*.

6.5 The square root method. The doublet rows of a Gauss-Doolittle solution may be replaced by a single row if the equations are symmetric. This is done by replacing the pairs of terms by their geometric means. Thus

$$s_{ij} = \sqrt{a_{ij}b_{ij}} = \sqrt{a_{ij}\frac{a_{ij}}{a_{ii}}} = \frac{a_{ij}}{\sqrt{a_{ii}}}$$

(1)

$$s_{ij\cdot(h)} = \sqrt{g_{ij\cdot(h)}b_{ij\cdot(h)}} = \sqrt{g_{ij\cdot(h)}\frac{g_{ij\cdot(h)}}{g_{ii\cdot(h)}}} = \frac{g_{ij\cdot(h)}}{\sqrt{g_{ii\cdot(h)}}}.$$

When $j = i$, (1) becomes

$$s_{ii} = \frac{a_{ii}}{\sqrt{a_{ii}}} = \sqrt{a_{ii}}$$

(2)

$$s_{ii\cdot(h)} = \frac{g_{ii\cdot(h)}}{\sqrt{g_{ii\cdot(h)}}} = \sqrt{g_{ii\cdot(h)}}.$$

The application of the usual pivotal condensation process leads to a method in which the diagonal terms are the square roots of $g_{ii\cdot(i-1)}$, the diagonal term of Table 6.4*j*.

It follows at once from (1) and (2) that

$$a_{ij} = s_{ii}s_{ij}$$

(3)

$$g_{ij\cdot(h)} = s_{ii\cdot(h)}s_{ij\cdot(h)}$$

and

$$b_{ij} = \frac{a_{ij}}{a_{ii}} = \frac{s_{ij}}{s_{ii}}$$

(4)

$$b_{ij\cdot(h)} = \frac{g_{ij\cdot(h)}}{g_{ii\cdot(h)}} = \frac{s_{ij\cdot(h)}}{s_{ii\cdot(h)}}$$

so that

(5) $$g_{ik\cdot(h)}b_{kj\cdot(h)} = s_{ik\cdot(h)}s_{kj\cdot(h)}.$$

Application of these formulas to (6.4.13) gives

$$g_{ij\cdot 1} = a_{ij} - s_{i1}s_{1j}$$

(6) $$g_{ij\cdot 12} = a_{ij} - s_{i1}s_{1j} - s_{i2\cdot 1}s_{2j\cdot 1}$$

$$g_{ij\cdot 123} = a_{ij} - s_{i1}s_{1j} - s_{i2\cdot 1}s_{2j\cdot 1} - s_{i3\cdot 12}s_{3j\cdot 12},$$

etc.

so that

(7) $$s_{ii\cdot(h)} = \sqrt{a_{ii} - s_{i1}s_{1i} - s_{i2\cdot 1}s_{2i\cdot 1} - \cdots - s_{ih\cdot(h-1)}s_{hi\cdot(h-1)}}$$

(8) $$s_{ij\cdot(h)} = \frac{a_{ij} - s_{i1}s_{1j} - s_{i2\cdot 1}s_{2j\cdot 1} - \cdots - s_{ih\cdot(h-1)}s_{hj\cdot(h-1)}}{s_{ii\cdot(h)}}.$$

Formula (7) describes an operational unit U_{13} whereas (8) is an operational unit of U_9 type.

The presentation of a square root method for non-symmetric matrices is given in Table 6.5*a*. The illustration is that of Table 6.4*h*.

The square root method is not superior, for computational purposes, to other compact forms of the method of single division in the non-symmetric case. In the symmetric case the values of $s_{ij\cdot(h)}$ are identical with those of $s_{ji\cdot(h)}$, so that all entries below the main diagonal of Table 6.5*a* may be omitted. The result is a solution that replaces the doublet rows of a Gauss-Doolittle solution with single rows. It has the general form of Table 6.4*j*. Some additional calculations are demanded such as extraction of square roots of diagonal terms, but its general feature of division as each term is computed, rather than the incorporation of the division in the next calculation, seems to make it preferable to the method of Table 6.4*j*. The method is based on the formulas

(9) $$s_{ii\cdot(h)} = \sqrt{a_{ii} - {s_{1i}}^2 - {s_{2i\cdot 1}}^2 - \cdots - {s_{hi\cdot(h-1)}}^2}$$

(10) $$s_{ij\cdot(h)} = \frac{a_{ij} - s_{1i}s_{1j} - s_{2i\cdot 1}s_{2j\cdot 1} - \cdots - s_{hi\cdot(h-1)}s_{hj\cdot(h-1)}}{s_{ii\cdot h}},$$

which are obtained at once from (7) and (8).

TABLE 6.5a

SQUARE ROOT METHOD FOR NON-SYMMETRIC CASES

x_1	x_2	x_3	x_4	a_{i5}	Check	Sum
a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a'_{15}	a_{1T}
a_{21}	a_{22}	a_{23}	a_{24}	a_{25}	a'_{25}	a_{2T}
a_{31}	a_{32}	a_{33}	a_{34}	a_{35}	a'_{35}	a_{3T}
a_{41}	a_{42}	a_{43}	a_{44}	a_{45}	a'_{45}	a_{4T}
s_{11}	s_{12}	s_{13}	s_{14}	s_{15}	sum	s_{1T}
s_{21}	$s_{22\cdot1}$	$s_{23\cdot1}$	$s_{24\cdot1}$	$s_{25\cdot1}$	sum	$s_{2T\cdot1}$
s_{31}	$s_{32\cdot1}$	$s_{33\cdot12}$	$s_{34\cdot12}$	$s_{35\cdot12}$	sum	$s_{3T\cdot12}$
s_{41}	$s_{42\cdot1}$	$s_{43\cdot12}$	$s_{44\cdot123}$	$s_{45\cdot123}$	sum	$s_{4T\cdot123}$
$x_{15\cdot234}$	$x_{25\cdot134}$	$x_{35\cdot124}$	$x_{45\cdot123}$			

x_1	x_2	x_3	x_4	a_{i5}	Check	Sum
26	−10	15	32	23	23	86
19	45	−14	−8	57	57	99
−12	16	27	13	47	47	91
32	29	−35	28	−68	−68	−14
5.099	−1.961	2.942	6.276	4.511	16.867	16.866
3.726	7.232	−3.452	−4.340	5.558	4.998	5.000
−2.353	1.574	6.273	5.515	7.790	19.578	19.578
6.276	5.712	−5.380	6.563	−13.126	−6.563	−6.564
2.000	1.000	3.000	−2.000			

The advantages of the square root method over the Gauss-Doolittle method are that it is compact, that it requires less recording, and that it permits greater ease in the finding of the entries to be used. The method is especially useful in obtaining the inverse matrix and in solving problems in statistics. Its chief disadvantage is that it requires the square root of each diagonal term. This process can be accomplished very quickly with modern machine methods, with the use of auxiliary tables, or with a slide rule.

The method we use does not matter greatly if we are working with three or four equations, but, if the number of equations is greater than

TABLE 6.5*b*

SQUARE ROOT METHOD FOR THE SYMMETRIC CASE

x_1	x_2	x_3	x_4	a_{45}	Check	Sum
a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a'_{15}	a_{1T}
*	a_{22}	a_{23}	a_{24}	a_{25}	a'_{25}	a_{2T}
*	*	a_{33}	a_{34}	a_{35}	a'_{35}	a_{3T}
*	*	*	a_{44}	a_{45}	a'_{45}	a_{4T}
s_{11}	s_{12}	s_{13}	s_{14}	s_{15}	sum	s_{1T}
	$s_{22\cdot1}$	$s_{23\cdot1}$	$s_{24\cdot1}$	$s_{25\cdot1}$	sum	$s_{2T\cdot1}$
		$s_{33\cdot12}$	$s_{34\cdot12}$	$s_{35\cdot12}$	sum	$s_{3T\cdot12}$
			$s_{44\cdot123}$	$s_{45\cdot123}$	sum	$s_{4T\cdot123}$
$x_{15\cdot234}$	$x_{25\cdot134}$	$x_{35\cdot124}$	$x_{45\cdot123}$			

x_1	x_2	x_3	x_4	a_{45}	Check	Sum
1.0000	0.4000	0.5000	0.6000	0.2000	0.19998	2.7000
*	1.0000	0.3000	0.4000	0.4000	0.40000	2.5000
*	*	1.0000	0.2000	0.6000	0.60006	2.6000
*	*	*	1.0000	0.8000	0.80000	3.0000
1.0000	0.4000	0.5000	0.6000	0.2000	2.7000	2.7000
	0.9165	0.1091	0.1746	0.3492	1.5494	1.5494
		0.8591	−0.1386	0.5377	1.2582	1.2582
			0.7683	0.9027	1.6710	1.6710
−0.9367	0.0601	0.8154	1.1749			

four, we should select a suitable method with care. The square root method demands $p = 4$ square roots and a slightly more involved operational unit (U_9 rather than U_3) in the back solution. On the other hand, the Gauss-Doolittle solution demands space for four more elimination equations and the transcription of fourteen additional entries, exclusive of checks.

It is true that some further approximation is introduced as a result of the square root operation, but the operation has a small relative error, see (2.12.3), as compared with the operations of multiplication and division, which are extensively featured. In most problems there is

no appreciable difference between the results derived through the use of the square root method and those through the use of the Gauss-Doolittle method if the incomplete numbers are carried to the same number of digits.

The square root method is especially applicable to situations in which $a_{ii\cdot(h)}$ is positive, as, for instance, in many of the problems of least squares and statistics. The technique may be used, even if $a_{ii\cdot(h)}$ is negative, although imaginary numbers may be thus introduced. Even in this case the method frequently features only positive and negative numbers when each of the entries in the row is of the form $(\ )i$ so that the product of any two of them is $(\ )i^2 = -(\ )$.

The square root method has been worked out independently by a number of different authors. Banachiewicz published it in 1938 in Poland [F.1, 2]. It has more recently been advocated by Dwyer [F.3], Duncan and Kenney [F.4], and Laderman [F.5] as an excellent method for studying least squares, correlation, and regression problems. Articles [F.6, 7, 8, 9] have traced the method to the earlier work of Cholesky and Schur. It is doubtful that such a simple method was not considered by authors prior to Cholesky. Before the days of computing machines, the method would probably have been discarded as impractical by any author who considered it.

REFERENCES

A. 1. E. V. Huntington, "Curve fitting by the method of least squares," *Handbook of Mathematical Statistics*, Houghton Mifflin Co., Boston, 1924, pp. 66–67.
2. C. R. Brolyer, "Another short method for computing β weights and resulting multiple R on a computing machine," *Journal of General Psychology*, **8**, 278–281 (1933).
3. Henry Chauncey, "A method for solving simultaneous equations, with particular reference to multiple correlation problems," *Harvard Educational Review*, **9**, 63–68 (1939).

B. Truman L. Kelley, "Partial and multiple correlation," *Handbook of Mathematical Statistics*, Houghton Mifflin Co., Boston, 1924, pp. 139–146. See p. 142.

C. 1. C. F. Gauss, "Supplementum theoriae combinationis observationum erroribus minimis obnoxiae," *Werke*, **15** (1873).
2. M. H. Doolittle, "Method employed in the solution of normal equations and the adjustment of a triangulation," *U. S. Coast and Geodetic Survey Report*, 1878, pp. 115–120.
3. F. V. Waugh, "A simplified method of determining multiple regression constants," *Journal of American Statistical Association*, **30**, 694–700 (1935).
4. P. S. Dwyer, "The solution of simultaneous equations," *Psychometrika*, **6**, 101–129 (1941).
5. P. S. Dwyer, "The Doolittle technique," *Annals of Mathematical Statistics*, **12**, 449–458 (1941).

6. A. C. Aitken, "Studies in practical mathematics. I. The evaluation, with applications, of a certain triple product matrix," *Proceedings Royal Society, Edinburgh*, **57**, 172–181 (1937).
7. T. Banachiewicz, "Études d'analyse pratique," *Cracow Observatory Reprint* 22, University of Cracow, 1938.
8. P. D. Crout, "A short method for evaluating determinants and solving systems of linear equations with real or complex coefficients," *Marchant Methods* MM-182, September 1941, Marchant Calculating Machine Co., Oakland, Calif.

D. 1. P. S. Dwyer, "A matrix presentation of least squares and correlation theory with matrix justification of improved methods of solution," *Annals of Mathematical Statistics*, **15**, 82–89 (1944).
2. H. H. Goldstine and J. von Neumann, "Numerical inverting of matrices of high order," *Bulletin of American Mathematical Society*, **53**, 1021–1099 (1947). See section 4.3.

E. F. E. Satterthwaite, "Error control in matrix calculation," *Annals of Mathematical Statistics*, **15**, pp. 373–387 (1944). See section 5.

F. 1. T. Banachiewicz, "Principes d'une nouvelle technique de la méthode des moindres carrés; Méthode de résolution numérique des équations linéaires, du calcul des déterminants et des inverses, et de réduction des formes quadratiques," Akademija Umiejetnosci, Krakow, *Wydzial Matematyczno-przyrodniczy, Bull. Intern., Sci. Math.*, 1938, pp. 134–135, 393–404.
2. T. Banachiewicz, "An outline of the cracovian algorithm of the method of least squares," *Astronomical Journal*, **50**, 38–41 (1942).
3. P. S. Dwyer, "The square root method and its use in correlation and regression," *Journal of the American Statistical Association*, **40**, 493–503 (1945).
4. D. B. Duncan and J. F. Kenney, *On the Solution of Normal Equations and Related Topics*, Edwards Brothers, Ann Arbor, Mich., 1946.
5. Jack Laderman, "The square root method for solving simultaneous linear equations," *Mathematical Tables and Other Aids to Computation*, **3**, 13–16 (1948).
6. L. Fox, H. D. Huskey, and J. H. Wilkinson, "Notes on the solution of algebraic linear simultaneous equations," *Quarterly Journal of Mechanics and Applied Mathematics*, **1**, 149–173 (1948).
7. A. M. Turing, "Rounding-off errors in matrix processes," *Quarterly Journal of Mechanics and Applied Mathematics*, **1**, 287–308 (1948).
8. M. Herzberger, "The normal equations of the method of least squares and their solution," *Quarterly of Applied Mathematics*, **7**, 217–223 (1949).
9. A. S. Householder, "Some numerical methods for solving systems of linear equations," *The American Mathematical Monthly*, **57**, 453–459 (1950).

EXERCISES

1. Solve exercise 4.10 by the method of row division.
2. Solve exercise 5.5 by the abbreviated method of row division.
3. Solve

$$1.000x_1 + 0.313x_2 + 0.280x_3 = 0.495$$

$$0.313x_1 + 1.000x_2 + 0.652x_3 = 0.650$$

$$0.280x_1 + 0.652x_2 + 1.000x_3 = 0.803$$

by the method of diagonal division.

4. Solve exercise 4.10 by the compact form of the method of single division.
5. Solve exercise 5.5 by the compact form of the method of single division.
6. Solve exercise 5.6 by the compact form of the method of single division.
7. Solve exercise 6.3 by the compact form of the method of single division.
8. Solve exercise 6.3 by the abbreviated Gauss-Doolittle method.
9. Solve exercise 4.10 by the generalized Gauss-Doolittle method.
10. Solve exercise 6.3 by the square root method.
11. Solve the symmetric equations:

$$\begin{aligned}
1.000x_1 + 0.313x_2 + 0.280x_3 + 0.182x_4 + 0.166x_5 &= 0.495 \\
* \quad + 1.000x_2 + 0.652x_3 + 0.554x_4 + 0.615x_5 &= 0.650 \\
* \quad + \quad * \quad + 1.000x_3 + 0.747x_4 + 0.693x_5 &= 0.803 \\
* \quad + \quad * \quad + \quad * \quad + 1.000x_4 + 0.774x_5 &= 0.804 \\
* \quad + \quad * \quad + \quad * \quad + \quad * \quad + 1.000x_5 &= 0.812
\end{aligned}$$

(*a*) By the compact method of single division.
(*b*) By the abbreviated Doolittle method.
(*c*) By the square root method.

CHAPTER 7

Relations between the Coefficients

7.1 Introduction. Several different types of coefficients appearing in the elimination equations obtained in the direct reduction of section 4.1.1 are introduced in Chapters 4 to 6. In particular the coefficients are $m_{ij\cdot(h)}$, $d_{ij\cdot(h)}$, $b_{ij\cdot(h)}$, $g_{ij\cdot(h)}$, and $s_{ij\cdot(h)}$. We might also mention in this connection the coefficients of the method of row division, but this method does not appear to be sufficiently useful to justify its further treatment here.

The notation for each of these coefficients and the definition (or the computational procedure associated with the definition) of each should be kept in mind as reference is frequently made to these methods in the remaining chapters. Thus $d_{34\cdot12}$ refers to a specified coefficient in one of the elimination equations obtained with the use of the method of multiplication and subtraction with (exact) division.

We improve the notation somewhat by defining

$$m_{ij} = a_{ij}, \qquad d_{ij} = a_{ij}, \qquad g_{ij} = a_{ij} \tag{1}$$

so that each set of equations can be written in more consistent notation.

This chapter examines the relations existing between these different types of coefficients. As a matter of fact, the coefficients of any system can be derived from the known coefficients of any other system.

Each formula of this chapter is an exact formula and does not involve approximations, even though divisions or square roots are indicated. The approximation enters only when we substitute in the algebraic formula and use digital numbers as approximations.

7.2 Relations previously indicated. Some of the results appropriate to this chapter appear, either explicitly or implicitly, in earlier chapters. Explicit formulas for $d_{ij\cdot(h)}$ in terms of m's, and for $m_{ij\cdot(h)}$ in terms of d's, are given in section 5.4, and the values of $b_{ij\cdot(h)}$ are given in terms of m's and d's in (6.3.13). Formulas (6.5.3) and (6.5.4) give the values of $g_{ij\cdot(h)}$ and $b_{ij\cdot(h)}$ in terms of s's. The elimination procedures of the last three chapters lead to a series of identities:

$$(1)\qquad \frac{m_{ij\cdot(h)}}{m_{ii\cdot(h)}} = \frac{d_{ij\cdot(h)}}{d_{ii\cdot(h)}} = \frac{b_{ij\cdot(h)}}{b_{ii\cdot(h)}} = \frac{g_{ij\cdot(h)}}{g_{ii\cdot(h)}} = \frac{s_{ij\cdot(h)}}{s_{ii\cdot(h)}} = x_{ij\cdot(h)}$$

with $b_{ii\cdot(h)} = 1$.

This series of identities gives at once the formulas for expressing $b_{ij\cdot(h)}$ in terms of the m's, d's, g's, and s's. The series also gives such identities as

$$(2)\qquad \begin{aligned} \frac{m_{ij\cdot(h)}}{b_{ij\cdot(h)}} &= m_{ii\cdot(h)} \\ \frac{d_{ij\cdot(h)}}{b_{ij\cdot(h)}} &= d_{ii\cdot(h)} \\ \frac{g_{ij\cdot(h)}}{b_{ij\cdot(h)}} &= g_{ii\cdot(h)} \\ \frac{s_{ij\cdot(h)}}{b_{ij\cdot(h)}} &= s_{ii\cdot(h)}. \end{aligned}$$

As was stated above and with the modification of (7.1.1), it is possible to derive formulas that give the coefficients of one system in terms of the coefficients of any other system except the b's. Formulas of this type are given in sections 5.4 and 7.3. Since this book is devoted to computational techniques and their justification, we concentrate on the selection of those formulas that are essential to this purpose. Section 5.4 exhibits formulas for transferring m's to d's and vice versa, whereas the formulas of section 6.5 enable us to transfer from g's to s's and from s's to g's. Formulas for d's in terms of g's, and for g's in terms of d's are necessary to complete the link. These formulas are provided in the next sections.

7.3 The values of $d_{ij\cdot(h)}$ in terms of g's.

$$d_{ij\cdot 1} = a_{11}a_{ij} - a_{i1}a_{1j} = a_{11}(a_{ij} - a_{i1}b_{1j}) = g_{11}g_{ij\cdot 1}$$

$$\begin{aligned} d_{ij\cdot 12} &= \frac{d_{22\cdot 1}d_{ij\cdot 1} - d_{i2\cdot 1}d_{2j\cdot 1}}{a_{11}} \\ &= \frac{a_{11}{}^2(g_{22\cdot 1}g_{ij\cdot 1} - g_{i2\cdot 1}g_{2j\cdot 1})}{a_{11}} = g_{11}g_{22\cdot 1}g_{ij\cdot 12}. \end{aligned}$$

$$\begin{aligned} d_{ij\cdot 123} &= \frac{d_{33\cdot 12}d_{ij\cdot 12} - d_{i3\cdot 12}d_{3j\cdot 12}}{d_{22\cdot 1}} \\ &= \frac{g_{11}{}^2g_{22\cdot 1}{}^2(g_{33\cdot 12}g_{ij\cdot 12} - g_{i3\cdot 12}g_{3j\cdot 12})}{g_{11}g_{22\cdot 1}} \\ &= g_{11}g_{22\cdot 1}g_{33\cdot 12}g_{ij\cdot 123}. \end{aligned}$$

It can be shown by mathematical induction that

$$d_{ij\cdot(h)} = g_{11}g_{22\cdot1}g_{33\cdot12}\cdots g_{hh\cdot(h-1)}g_{ij\cdot(h)}. \tag{1}$$

As a special case of (1) we have

$$d_{hh\cdot(h-1)} = g_{11}g_{22\cdot1}g_{33\cdot12}\cdots g_{hh\cdot(h-1)}. \tag{2}$$

7.4 The values of $g_{ij\cdot(h)}$ in terms of the d's. Substitution of (7.3.2) in (7.3.1) yields at once

$$d_{ij\cdot(h)} = d_{hh\cdot(h-1)}g_{ij\cdot(h)} \tag{1}$$

from which we obtain

$$g_{ij\cdot(h)} = \frac{d_{ij\cdot(h)}}{d_{hh\cdot(h-1)}}. \tag{2}$$

This formula shows how a solution in terms of d's may be used to give a solution in terms of g's by dividing by the preceding pivot. As an example, $d_{34\cdot12} = 0.5120$ in Table 5.8a, divided by $d_{22\cdot1} = 0.8400$, yields 0.6095, which is the value of $g_{34\cdot12}$ of Table 6.4a.

This formula may also appear in the form [A],

$$\frac{d_{ij\cdot(h)}}{g_{ij\cdot(h)}} = d_{hh\cdot(h-1)}. \tag{3}$$

A somewhat similar formula

$$\frac{d_{ij\cdot(h)}}{b_{ij\cdot(h)}} = d_{ii\cdot(h)} \tag{4}$$

is given in (7.2.2).

7.5 Additional formulas. The foregoing formulas are presented with a view to showing how any set of coefficients can be transformed directly to any other set of coefficients. The formulas selected are chosen with a view toward their use in later chapters and with a view toward minimizing the mathematical manipulation of formulas. Actually we could derive many identities, but this extensive development is hardly appropriate to our objective.

Since the d's are featured in connection with exact methods and the s's in connection with symmetric approximate methods in the later chapters, it seems wise to relate them. Application of (6.5.2) and (6.5.1) to (7.3.1) gives

$$d_{ij\cdot(h)} = s^2_{11}s^2_{22\cdot1}s^2_{33\cdot12}\cdots s_{ii\cdot(h)}s_{ij\cdot(h)} \tag{1}$$

with

$$d_{ii\cdot(h)} = s^2_{11}s^2_{22\cdot1}s^2_{33\cdot12}\cdots s^2_{ii\cdot(h)}. \tag{2}$$

It follows that

$$s_{ij\cdot(h)} = \frac{d_{ij\cdot(h)}}{\sqrt{d_{ii\cdot(h)}d_{hh\cdot(h-1)}}}. \tag{3}$$

This formula can be used in getting the values $s_{ij\cdot(h)}$ from the values of Chapter 5 (no matter whether symmetry is present or not) by taking the element and dividing it by the geometric mean of the leading element of the ith row and the pivot just before it.

REFERENCES

A. F. W. Waugh and P. S. Dwyer, "The compact computation of the inverse of a matrix," *Annals of Mathematical Statistics*, **16**, 259–271 (1945). See p. 267.

EXERCISES

1. Compute the value of $d_{33\cdot12}$ for exercise 6.4.
2. Compute the value of $d_{44\cdot123}$ for exercise 5.5.
3. Compute the value of $d_{44\cdot123}$ for exercise 5.6.
4. Compute the value of $d_{33\cdot12}$ for exercise 6.7.
5. Compute the value of $d_{33\cdot12}$ for exercise 6.8.
6. Compute the value of $d_{44\cdot123}$ for exercise 6.9.
7. Compute the value of $d_{33\cdot12}$ for exercise 6.10.
8. Compute the value of $d_{55\cdot1234}$ for exercise 6.11.
9. Eliminate in order x_1, x_2, x_3, etc., by the method of multiplication and subtraction with exact division for equations of exercise 6.11. Use (7.5.3) in computing some of the values of $s_{ij\cdot(h)}$. Check with the results of exercise 6.11.

CHAPTER 8

The Solution of Related and Associated Equations

8.1 Introduction. Sets of equations that may be obtained from a given set of equations such as (4.1.1) by deletion of one or more of the variables and equations may be said to be *related*. Thus

$$
\begin{aligned}
&a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = a_{15} \\
(1)\qquad &a_{21}x_1 + a_{22}x_2 + a_{23}x_3 = a_{25} \\
&a_{31}x_1 + a_{32}x_2 + a_{33}x_3 = a_{35}
\end{aligned}
$$

$$
\begin{aligned}
(2)\qquad &a_{11}x_1 + a_{12}x_2 = a_{15} \\
&a_{21}x_1 + a_{22}x_2 = a_{25}
\end{aligned}
$$

$$
\begin{aligned}
(3)\qquad &a_{22}x_2 + a_{23}x_3 = a_{25} \\
&a_{32}x_2 + a_{33}x_3 = a_{35}
\end{aligned}
$$

as well as (6.3.11) are all related. ✓

The term, related, may also apply to equations formed entirely from the left-hand coefficients of (4.1.1). We may say, for example, that the equations (6.3.12) and

$$
\begin{aligned}
(4)\qquad &a_{11}x_1 + a_{12}x_2 = a_{13} \\
&a_{21}x_1 + a_{22}x_2 = a_{23}
\end{aligned}
$$

are also related to the equations (4.1.2).

We next note the useful fact that the solution of many sets of related equations can be obtained from the same forward solution by the simple device of carrying through a back solution for each set of related equations. This idea, which was emphasized by Kurtz in connection with

the Doolittle method [A.1], can be applied to any of the methods outlined in Chapters 4 to 6. For purposes of illustration, application is made in this chapter to the four-variable problem used earlier (Table 4.8*a* and [A.2]) and to a six-variable problem that the author has used previously to illustrate this point [A.3].

A precise notation for each solution, such as that introduced in section 6.3, is necessary if we are to distinguish the different values of x_i that result from the use of different sets of related equations.

8.2 The solution of related equations. As suggested in the last section, a single forward solution may be used to obtain the solutions of

TABLE 8.2*a*

SOLUTION OF RELATED EQUATIONS

x_1	x_2	x_3	x_4	a_{45}	Check	Sum
a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a'_{15}	a_{1T}
a_{21}	a_{22}	a_{23}	a_{24}	a_{25}	a'_{25}	a_{2T}
a_{31}	a_{32}	a_{33}	a_{34}	a_{35}	a'_{35}	a_{3T}
a_{41}	a_{42}	a_{43}	a_{44}	a_{45}	a'_{45}	a_{4T}
a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	sum	a_{1T}
1	b_{12}	b_{13}	b_{14}	b_{15}	sum	b_{1T}
	$g_{22\cdot1}$	$g_{23\cdot1}$	$g_{24\cdot1}$	$g_{25\cdot1}$	sum	$g_{2T\cdot1}$
	1	$b_{23\cdot1}$	$b_{24\cdot1}$	$b_{25\cdot1}$	sum	$b_{2T\cdot1}$
		$g_{33\cdot12}$	$g_{34\cdot12}$	$g_{35\cdot12}$	sum	$g_{3T\cdot12}$
		1	$b_{34\cdot12}$	$b_{35\cdot12}$	sum	$b_{3T\cdot12}$
			$g_{44\cdot123}$	$g_{45\cdot123}$	sum	$g_{4T\cdot123}$
			1	$b_{45\cdot123}$	sum	$b_{4T\cdot123}$
$b_{15\cdot234}$	$b_{25\cdot134}$	$b_{35\cdot124}$	$b_{45\cdot123}$	-1		
$b_{15\cdot23}$	$b_{25\cdot13}$	$b_{35\cdot12}$		-1		
$b_{15\cdot2}$	$b_{25\cdot1}$			-1		
b_{15}				-1		

many related equations. The Gauss-Doolittle method might be used. The values of the solutions of (4.1.2) and (8.1.1) and (8.1.2) are given at the bottom of Table 8.2*a*. The forward solution corresponds to that of Table 6.4*g*, and the first row of the back solution indicates the solution of (4.1.2). The explicit algebraic identities that are the basis

of the back solution of (4.1.2) now appear as

$$
\begin{aligned}
b_{45\cdot 123} &= b_{45\cdot 123} \\
b_{35\cdot 124} &= b_{35\cdot 12} - b_{34\cdot 12}b_{45\cdot 123} \\
b_{25\cdot 134} &= b_{25\cdot 1} - b_{24\cdot 1}b_{45\cdot 123} - b_{23\cdot 1}b_{35\cdot 124} \\
b_{15\cdot 234} &= b_{15} - b_{14}b_{45\cdot 123} - b_{13}b_{35\cdot 124} - b_{12}b_{25\cdot 134}.
\end{aligned} \tag{1}
$$

These algebraic identities describe the back solution in a precise notation.

The solution of (8.1.1) is obtained by ignoring, in the back solution, all the terms of Table 8.2*a* that have 4's as subscripts. Thus

$$
\begin{aligned}
b_{35\cdot 12} &= b_{35\cdot 12} \\
b_{25\cdot 13} &= b_{25\cdot 1} - b_{23\cdot 1}b_{35\cdot 12} \\
b_{15\cdot 23} &= b_{15} - b_{13}b_{35\cdot 12} - b_{12}b_{25\cdot 13}.
\end{aligned} \tag{2}
$$

In a similar fashion the solutions of (8.1.2) are

$$
\begin{aligned}
b_{25\cdot 1} &= b_{25\cdot 1} \\
b_{15\cdot 2} &= b_{15} - b_{12}b_{25\cdot 1}.
\end{aligned} \tag{3}
$$

The formal proof that these are identities results from the fact that they are the values of x_1, x_2, x_3, and x_4 and that they must satisfy all the elimination equations as well as the original equations. The value -1 is inserted under the values of the a_{i5} so that each identity appears in the form $\Sigma uv = 0$, where u is the element in the row of the forward solution and v is the corresponding element in the row of the solution. Using this scheme of writing the identities we could write at once a large number of identities, some of which are trivial, from the Table 8.2*a*. Care must be taken to use only equations from the two groups in which the secondary subscripts of the b's are equal to the first primary subscripts of the g's. Thus

$$b_{15\cdot 23}g_{41} + b_{25\cdot 13}g_{42} + b_{35\cdot 12}g_{43} - a_{45} \neq 0$$

since the b's are solutions of equations with a_{i5}, and not a_{i4}, as the right-hand term.

It is probably wise to write the coefficients of lower order first if we are to calculate all these regression coefficients. A numerical illustration of this is presented in Table 8.2*b*, where the illustration of Table 6.4*g* is used.

TABLE 8.2b

SOLUTION OF RELATED EQUATIONS—ILLUSTRATION

x_1	x_2	x_3	x_4	a_{i5}	Check	Sum
1.0000	0.4000	0.5000	0.6000	0.2000		2.7000
*	1.0000	0.3000	0.4000	0.4000		2.5000
*	*	1.0000	0.2000	0.6000		2.6000
*	*	*	1.0000	0.8000		3.0000
1.0000	0.4000	0.5000	0.6000	0.2000	2.70000	2.7000
1.0000	0.4000	0.5000	0.6000	0.2000	2.70000	2.7000
	0.8400	0.1000	0.1600	0.3200	1.4200	1.4200
	1.0000	0.1190	0.1905	0.3810	1.6905	1.6905
		0.7381	−0.1190	0.4619	1.0810	1.0810
		1.0000	−0.1612	0.6258	1.4646	1.4646
			0.5903	0.6935	1.2838	1.2838
			1.0000	1.1748	2.1748	2.1748
0.2000						
0.0476	0.3810					
−0.2355	0.3065	0.6258				
−0.9366	0.0602	0.8152	1.1748			

A solution to (6.3.12) may also be obtained from the forward solution of (4.1.2). In this case all entries containing 5's as subscripts should be ignored. We get at once

$$
\begin{aligned}
b_{34\cdot 12} &= b_{34\cdot 12} \\
b_{24\cdot 13} &= b_{24\cdot 1} - b_{23\cdot 1}b_{34\cdot 12} \\
b_{14\cdot 23} &= b_{14} - b_{13}b_{34\cdot 12} - b_{12}b_{24\cdot 31},
\end{aligned} \tag{4}
$$

and in a similar fashion

$$
\begin{aligned}
b_{23\cdot 1} &= b_{23\cdot 1} \\
b_{13\cdot 2} &= b_{13} - b_{12}b_{23\cdot 1}
\end{aligned} \tag{5}
$$

is the solution of (8.1.4).

The solutions of the fifteen sets of related equations that can be obtained from a single forward solution when $p = 5$ is shown in Table 8.2c,

TABLE 8.2*c*

SOLUTION OF MANY SETS OF RELATED EQUATIONS

x_1	x_2	x_3	x_4	x_5	
a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a_{16}
*	a_{22}	a_{23}	a_{24}	a_{25}	a_{26}
*	*	a_{33}	a_{34}	a_{35}	a_{36}
*	*	*	a_{44}	a_{45}	a_{46}
*	*	*	*	a_{55}	a_{56}
s_{11}	s_{12}	s_{13}	s_{14}	s_{15}	s_{16}
	$s_{22\cdot1}$	$s_{23\cdot1}$	$s_{24\cdot1}$	$s_{25\cdot1}$	$s_{26\cdot1}$
		$s_{33\cdot12}$	$s_{34\cdot12}$	$s_{35\cdot12}$	$s_{36\cdot12}$
			$s_{44\cdot123}$	$s_{45\cdot123}$	$s_{46\cdot123}$
				$s_{55\cdot1234}$	$s_{56\cdot1234}$
b_{15}					-1
$b_{16\cdot2}$	$b_{26\cdot1}$				-1
$b_{16\cdot23}$	$b_{26\cdot13}$	$b_{36\cdot12}$			-1
$b_{16\cdot234}$	$b_{26\cdot134}$	$b_{36\cdot124}$	$b_{46\cdot123}$		-1
$b_{16\cdot2345}$	$b_{26\cdot1345}$	$b_{36\cdot1245}$	$b_{46\cdot1235}$	$b_{56\cdot1234}$	-1
b_{15}				-1	
$b_{15\cdot2}$	$b_{25\cdot1}$			-1	
$b_{15\cdot23}$	$b_{25\cdot13}$	$b_{35\cdot12}$		-1	
$b_{15\cdot234}$	$b_{25\cdot134}$	$b_{35\cdot124}$	$b_{45\cdot123}$	-1	
b_{14}			-1		
$b_{14\cdot2}$	$b_{24\cdot1}$		-1		
$b_{14\cdot23}$	$b_{24\cdot13}$	$b_{34\cdot12}$	-1		
b_{13}		-1			
$b_{13\cdot2}$	$b_{23\cdot1}$	-1			
b_{12}	-1				

where the method of solution is the square root method. The -1 is used to indicate the column containing the coefficients of the right side of the equation. The row sum check is omitted for simplicity of presentation although it is probably wise to use it in a computational problem of this magnitude. An illustration used previously [A] is presented in Table 8.2*d*.

TABLE 8.2*d*

SOLUTION OF MANY SETS OF RELATED EQUATIONS—CARVER DATA

x_1	x_2	x_3	x_4	x_5	
1.000	0.313	0.280	0.182	0.166	0.495
*	1.000	0.652	0.554	0.615	0.650
*	*	1.000	0.747	0.693	0.803
*	*	*	1.000	0.774	0.804
*	*	*	*	1.000	0.812
1.000	0.313	0.280	0.182	0.166	0.495
	0.950	0.594	0.523	0.593	0.521
		0.754	0.511	0.390	0.471
			0.657	0.357	0.306
				0.584	0.219
0.495					−1
0.323	0.548				−1
0.271	0.158	0.625			−1
0.293	0.099	0.309	0.466		−1
0.310	0.012	0.253	0.262	0.375	−1
0.166				−1	
−0.029	0.624			−1	
−0.073	0.301	0.517		−1	
−0.047	0.232	0.149	0.543	−1	
0.182			−1		
0.010	0.551		−1		
−0.048	0.127	0.678	−1		
0.280		−1			
0.084	0.625	−1			
0.313	−1				

8.3 Relations between the solutions of sets of related equations. The formulas (8.2.1), (8.2.2), and (8.2.3) are special cases of more general formulas. The second formulas of each set are special cases of the formula

$$b_{ij\cdot\ \ldots\ h} = b_{ij\cdot\ \ldots} - b_{ih\cdot\ \ldots}\, b_{hj\cdot\ \ldots} \tag{1}$$

where $\cdots$ indicate fixed secondary subscripts. These indicate variables that have been eliminated in the condensation process. Letting e represent these eliminated variables, we have the symbolic form of (1)

$$b_{ij\cdot eh} = b_{ij\cdot e} - b_{ih\cdot e}b_{hj\cdot ei}. \tag{2}$$

It is shown later (10.9.5) that this is a true identity.

Similarly the third formulas are special cases of the symbolic formula

$$b_{ij\cdot ehl} = b_{ij\cdot e} - b_{il\cdot e}b_{lj\cdot ehi} - b_{ih\cdot e}b_{hj\cdot eli}. \tag{3}$$

It is to be noted that continued application of (1) yields

$$\begin{aligned} b_{ij\cdot h} &= b_{ij} - b_{ih}b_{hj\cdot i} \\ b_{ij\cdot hl} &= b_{ij\cdot h} - b_{il\cdot h}b_{lj\cdot hi} \\ &= b_{ij} - b_{ih}b_{hj\cdot i} - b_{il\cdot h}b_{lj\cdot hi} \\ b_{ij\cdot hlm} &= b_{ij} - b_{ih}b_{hj\cdot i} - b_{il\cdot h}b_{lj\cdot hi} - b_{im\cdot hl}b_{mj\cdot hli} \\ &\text{etc.} \end{aligned} \tag{4}$$

8.4 The solution of associated equations. Sometimes we wish to solve systems of equations that have identical coefficients on the left, but differ on the right side. Such sets of equations may be said to be *associated*.

The solution may be carried out in different ways. If the number of sets of these associated equations is not so very large, it may be wise to list each additional set and to carry through the forward solution as before. The row sum check now utilizes all the entries in the row. The back solution is carried out separately for each set of associated equations. The method is illustrated in Tables 8.4*a* and 8.4*b*, where three sets of equations having the same left-hand coefficients are worked simultaneously with the Gauss-Doolittle solution and the compact method of determinants. Other methods could be used.

If the number of associated equations is large, the solution should be obtained in terms of general expressions on the right. The numerical values that identify a given set may be substituted in the answer. For

* The existence of this symbolic formula was called to my attention by Dr. Paul Boschan.

TABLE 8.4a

SIMULTANEOUS SOLUTION, GAUSS-DOOLITTLE (WITH CHECK)

x_1	x_2	x_3	x_4				Check	Sum
1.0000	0.4000	0.5000	0.6000	0.2000	0.4000	0.8000		3.9000
*	1.0000	0.3000	0.4000	0.4000	0.5000	0.6000		3.6000
*	*	1.0000	0.2000	0.6000	0.6000	0.4000		3.6000
*	*	*	1.0000	0.8000	0.7000	0.2000		3.9000
1.0000	0.4000	0.5000	0.6000	0.2000	0.4000	0.8000	3.9000	3.9000
1.0000	0.4000	0.5000	0.6000	0.2000	0.4000	0.8000	3.9000	3.9000
	0.8400	0.1000	0.1600	0.3200	0.3400	0.2800	2.0400	2.0400
	1.0000	0.1190	0.1905	0.3810	0.4048	0.3333	2.4286	2.4286
		0.7381	−0.1190	0.4619	0.3595	−0.0333	1.4072	1.4071
		1.0000	−0.1612	0.6258	0.4871	−0.0451	1.9066	1.9064
			0.5903	0.6935	0.4532	−0.3387	1.3983	1.3983
			1.0000	1.1748	0.7677	−0.5738	2.3687	2.3688
−0.9366	0.0602	0.8152	1.1748					
−0.4404	0.1859	0.6109	0.7677					
1.0295	0.4590	−0.1376	−0.5738					

example, (4.1.2) can be written in the form

$$
\begin{aligned}
a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + a_{14}x_4 &= 1a_{15} + 0a_{25} + 0a_{35} + 0a_{45} \\
a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + a_{24}x_4 &= 0a_{15} + 1a_{25} + 0a_{35} + 0a_{45} \\
a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + a_{34}x_4 &= 0a_{15} + 0a_{25} + 1a_{35} + 0a_{45} \\
a_{41}x_1 + a_{42}x_2 + a_{43}x_3 + a_{44}x_4 &= 0a_{15} + 0a_{25} + 0a_{35} + 1a_{45}.
\end{aligned} \tag{1}
$$

Each of the results x_1, x_2, x_3, x_4 can then be obtained as

$$( )a_{15} + ( )a_{25} + ( )a_{35} + ( )a_{45}.$$

TABLE 8.4*b*

SIMULTANEOUS EQUATIONS—COMPACT

x_1	x_2	x_3	x_4				Check	Sum
1.0000	0.4000	0.5000	0.6000	0.2000	0.4000	0.8000		3.9000
*	1.0000	0.3000	0.4000	0.4000	0.5000	0.6000		3.6000
*	*	1.0000	0.2000	0.6000	0.6000	0.4000		3.6000
*	*	*	1.0000	0.8000	0.7000	0.2000		3.9000
1.0000	0.4000	0.5000	0.6000	0.2000	0.4000	0.8000	3.9000	3.9000
	0.8400	0.1000	0.1600	0.3200	0.3400	0.2800	2.0400	2.0400
		0.6200	−0.1000	0.3880	0.3020	−0.0280	1.1820	1.1820
			0.3660	0.4300	0.2810	−0.2100	0.8670	0.8670
−0.9366	0.0601	0.8153	1.1749					
−0.4404	0.1858	0.6109	0.7678					
1.0295	0.4590	−0.1377	−0.5738					

A synthetic form may be used for numerical work. Thus the equations might appear as shown below.

x_1	x_2	x_3	x_4	a_{15}	a_{25}	a_{35}	a_{45}
a_{11}	a_{12}	a_{13}	a_{14}	1	0	0	0
a_{21}	a_{22}	a_{23}	a_{24}	0	1	0	0
a_{31}	a_{32}	a_{33}	a_{34}	0	0	1	0
a_{41}	a_{42}	a_{43}	a_{44}	0	0	0	1

Illustrations using the method of Chapter 4 and the square root method are presented in Tables 8.4*c* and 8.4*d*.

A final verification consists in showing that each of the equations is satisfied by each of the four expressions. These values can then be used in obtaining the answers to problems involving the associated equations. For the example above we have from a symmetric form of the results of Table 8.4*c*.

TABLE 8.4c

METHOD OF CHAPTER 4

x_1	x_2	x_3	x_4	a_{15}	a_{25}	a_{35}	a_{45}	Sum
1.0000	0.4000	0.5000	0.6000	1	0	0	0	3.5000
*	1.0000	0.3000	0.4000	0	1	0	0	3.1000
*	*	1.0000	0.2000	0	0	1	0	3.0000
*	*	*	1.0000	0	0	0	1	3.2000
1.0000	0.4000	0.5000	0.6000	1.0000	0	0	0	3.5000
	0.8400	0.1000	0.1600	−0.4000	1.0000	0	0	1.7000
		0.6200	−0.1000	−0.3800	−0.1000	0.8400	0	0.8800
			0.30744	−0.3108	−0.1092	0.0840	0.5208	0.49224
2.0711	−0.1913	−0.7760	−1.0109					
−0.1913	1.2842	−0.2186	−0.3552					
−0.7759	−0.2186	1.3989	0.2732					
−1.0109	−0.3552	0.2732	1.6940					

TABLE 8.4d

SQUARE ROOT METHOD

x_1	x_2	x_3	x_4	a_{15}	a_{25}	a_{35}	a_{45}	Check	Sum
1.0000	0.4000	0.5000	0.6000	1	0	0	0		3.5000
*	1.0000	0.3000	0.4000	0	1	0	0		3.1000
*	*	1.0000	0.2000	0	0	1	0		3.0000
*	*	*	1.0000	0	0	0	1		3.2000
1.0000	0.4000	0.5000	0.6000	1	0	0	0	3.5000	3.5000
	0.9165	0.1091	0.1746	−0.4364	1.0911	0	0	1.8549	1.8549
		0.8591	−0.1386	−0.5206	−0.1386	1.1640	0	1.2193	1.2195
			0.7683	−0.7768	−0.2730	0.2100	1.3016	1.2301	1.2300
2.0712	−0.1912	−0.7761	−1.0111						
−0.1912	1.2842	−0.2187	−0.3553						
0.7760	−0.2186	1.3990	0.2733						
−1.0110	−0.3553	0.2733	1.6941						

$$
\begin{aligned}
x_1 &= 2.0711a_{15} - 0.1913a_{25} - 0.7759a_{35} - 1.0109a_{45} \\
x_2 &= -0.1913a_{15} + 1.2842a_{25} - 0.2186a_{35} - 0.3552a_{45} \\
x_3 &= -0.7759a_{15} - 0.2186a_{25} + 1.3989a_{35} + 0.2732a_{45} \\
x_4 &= -1.0109a_{15} - 0.3552a_{25} + 0.2732a_{35} + 1.6940a_{45}.
\end{aligned} \tag{2}
$$

When $a_{15} = 0.2000$, $a_{25} = 0.4000$, $a_{35} = 0.6000$, $a_{45} = 0.8000$, we have

$$x_1 = -0.9366, \qquad x_2 = 0.0601, \qquad x_3 = 0.8153, \qquad x_4 = 1.1749$$

as indicated in Table 8.4*b*. The other solutions of Table 8.4*b* are obtained similarly.

For further discussion of the solution of associated equations the reader is referred to the treatment of inverse matrices in Chapter 13.

REFERENCES

A. 1. A. K. Kurtz, "The use of the Doolittle method in obtaining related multiple correlation coefficients," *Psychometrika*, **1**, 45–51 (1936).

2. P. S. Dwyer, "The solution of simultaneous equations," *Psychometrika*, **6**, 101–129 (1941). See p. 118.

3. P. S. Dwyer, "Recent developments in correction technique," *Journal of the American Statistical Association*, **37**, 441–460 (1942). See p. 449.

EXERCISES

1. Solve the four equations of exercise 6.11 obtained by deleting the last equation and also the x_5 terms. From the same forward solution carry out the back solution when the x_4 terms; the x_4 terms and the x_3 terms; the x_4 terms, the x_3 terms, and the x_2 terms are also deleted. Check your results with those of Table 8.2*d*.

2. Solve the equations of Table 4.13*a* with the value of a_{15} replaced by 86, 99, 90, −14, respectively. Use the method of multiplication and subtraction with exact division.

3. Solve the equations of exercise 4.10 with the right-hand terms replaced by 111, 67, 73, and 89, respectively.

4. Solve exercise 5.5 with the first row and second column deleted.

5. Solve the equations of exercise 6.11 with the square root method if the values on the right are replaced by 1, 0, 0, 0, 0; 0, 1, 0, 0, 0; etc., according to the method of Table 8.4*d*. Check the results of exercise 6.11.

CHAPTER 9

Introduction to Determinants

9.1 Introduction. This chapter presents a brief introduction to determinant theory for those readers not familiar with the subject. It gives an explanation of the concept and states the fundamental theorems that are the basis of the usual computational procedure. No attempt is made to derive the theory involved, since it can be found in texts on algebra, theory of equations, and determinants. Because of the rudimentary nature of this material, the reader familiar with determinants will probably wish to do no more than to glance through the following pages.

9.2 Definition of a determinant. Determinants are usually introduced as synthetic devices for recording the results of solutions of simultaneous linear equations, but the formal definition, as usually given, states that the determinant value is a certain function of given elements. The elements of the determinant are arranged in the form of a square matrix, and the value of the determinant is indicated by a vertical line placed on each side of the matrix. The symbol Δ (or sometimes D) is used to denote the determinant. The order of the determinant is the number of rows and columns of the square matrix. The definition of a determinant of order 2, for example, is

$$\Delta = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc. \tag{1}$$

For example,

$$\begin{vmatrix} 2 & 3 \\ 1 & -1 \end{vmatrix} = -5 \quad \text{and} \quad \begin{vmatrix} -1 & -4 \\ 2 & -3 \end{vmatrix} = 11.$$

In general

$$\Delta = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{21}a_{12} = m_{22\cdot 1} = d_{22\cdot 1}. \tag{2}$$

In a similar way the determinant of order 3 is defined as

$$(3) \qquad \Delta = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}$$

$$= a_{11}a_{22}a_{33} + a_{21}a_{32}a_{13} + a_{31}a_{12}a_{23} - a_{31}a_{22}a_{13} - a_{21}a_{12}a_{33} - a_{11}a_{32}a_{23}.$$

The expansions (2) and (3) indicate that the determinant is defined by forming all possible products that can be obtained by taking one element out of one column and one row, one out of another column and another row, etc., by multiplying these elements together, by affixing an appropriate sign, and then adding. The rule for determining the appropriate sign may be stated in different ways. One accepted method indicates that any term of the determinant should be written with the second subscripts of its elements in 1, 2, 3 order, as illustrated in (2) and (3). The algebraic sign is taken as positive if an even number of interchanges of first subscripts reduces them to 1, 2, 3 order and as negative if an odd number of interchanges is necessary. This rule can be applied to give the proper sign of any term of (2) and (3). For example, the term $a_{21}a_{12}$ has a minus sign since the first subscripts are in 2, 1 order and it takes one inversion to bring them into 1, 2 order. Similarly the term $a_{31}a_{12}a_{23}$ has a plus sign since the order 3, 1, 2 can be reduced to the 1, 2, 3 order by two inversions.

The rules specified above may be taken as the definition of a determinant of any order. For example, the determinant

$$(4) \qquad \Delta = \begin{vmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{vmatrix}$$

is defined to be the sum of the 4! (or 24 terms) resulting from application of the above rules.

9.3 Properties of determinants. From the definition we are able to prove various properties that are very useful in manipulation with determinants and that serve as the basis of the classical procedure for the numerical evaluation of determinants. Some of the important rules are:

(*a*) If corresponding rows and columns are interchanged, the value of the determinant is unchanged.

(*b*) If two rows (or columns) are interchanged, the sign of the determinant is unchanged.

(*c*) If two rows (or columns) are identical, the value of the determinant is zero.

(*d*) If every element of a row (or column) is multiplied by the same number, the determinant is multiplied by that number.

(*e*) If every element of a row (or column) is zero, the value of the determinant is zero.

(*f*) If to the elements of every row (or column) are added the corresponding elements of any other row (or column) each multiplied by the same arbitrary number, the value of the determinant is unchanged. Subtraction is also permissible since the arbitrary number may be negative.

9.4 Expansion of determinants. A determinant may be *expanded* in terms of the elements of any column (or those of any row). For example, the 24 terms of the expansion of (9.2.4) may be written as

$$\Delta = a_{11}\begin{vmatrix} a_{22} & a_{23} & a_{24} \\ a_{32} & a_{33} & a_{34} \\ a_{42} & a_{43} & a_{44} \end{vmatrix} - a_{21}\begin{vmatrix} a_{12} & a_{13} & a_{14} \\ a_{32} & a_{33} & a_{34} \\ a_{42} & a_{43} & a_{44} \end{vmatrix} \tag{1}$$
$$+a_{31}\begin{vmatrix} a_{12} & a_{13} & a_{14} \\ a_{22} & a_{23} & a_{24} \\ a_{42} & a_{43} & a_{44} \end{vmatrix} - a_{41}\begin{vmatrix} a_{12} & a_{13} & a_{14} \\ a_{22} & a_{23} & a_{24} \\ a_{32} & a_{33} & a_{34} \end{vmatrix}$$

where the expansion is in terms of the elements of the first column. The determinantal coefficients of $\pm a_{ij}$ in (1) are known as *minors* and are obtained by deleting the ith row and the jth column of (9.2.4). The signs are determined by the number of moves necessary to put a_{ij} in the leading position. An even number indicates a plus sign and an odd number a minus sign. The coefficients of the a_{ij} are known as *cofactors*. The expansion may be made in terms of the elements of any row, or those of any column, as long as the foregoing rules are observed. For example, the determinant

$$\begin{vmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \\ 0 & 2 & 1 \end{vmatrix} = 1\begin{vmatrix} 3 & 1 \\ 2 & 1 \end{vmatrix} - 2\begin{vmatrix} 2 & 3 \\ 2 & 1 \end{vmatrix} + 0\begin{vmatrix} 2 & 3 \\ 3 & 1 \end{vmatrix}$$

$$= 1(1) - 2(-4) + 0(-7) = 9$$

by expanding in terms of the elements of the first row.

It is conventional to use an exclusion notation in indicating determinants that are formed from a determinant Δ by deletion of rows and columns. Thus Δ_{ij} is the determinant formed by deletion of the ith row and the jth column. Similarly $\Delta_{ij\cdot kl}$ is formed by deleting the ith row and the jth column and the kth row and the lth column. Of course $\Delta_{ii\cdot jj} = \Delta_{jj\cdot ii}$ since both result from the deletion of the ith and jth rows and columns.

9.5 Use of determinants in solving equations. Although determinants have other uses, our chief interest here is in the application to the solution of simultaneous linear equations. It can be proved from the principles outlined above that the solution of (4.1.2) can be expressed as

$$(1) \qquad x_4 = x_{45\cdot 123} = b_{45\cdot 123} = \frac{\begin{vmatrix} a_{11} & a_{12} & a_{13} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{25} \\ a_{31} & a_{32} & a_{33} & a_{35} \\ a_{41} & a_{42} & a_{43} & a_{45} \end{vmatrix}}{\begin{vmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{vmatrix}}.$$

Similar determinantal expressions hold for

$$(2) \qquad \begin{aligned} x_3 &= x_{35\cdot 124} = b_{35\cdot 124} \\ x_2 &= x_{25\cdot 134} = b_{25\cdot 134} \\ x_1 &= x_{15\cdot 234} = b_{15\cdot 234}. \end{aligned}$$

The denominator determinant is the same in all cases whereas the a_{i5} column replaces the a_{i4} column, the a_{i3} column, the a_{i2} column, and the a_{i1} column in turn to form the successive numerators of x_4, x_3, x_2, x_1. The above illustrates the theorem known as Cramer's rule.

If we consider the determinant

$$(3) \qquad \Delta = \begin{vmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} \\ a_{51} & a_{52} & a_{53} & a_{54} & a_{55} \end{vmatrix}.$$

with the a_{i5} terms either specified or unspecified, the values of the solution of (4.1.2) may be indicated as

$$
\begin{aligned}
x_{45\cdot 123} &= b_{45\cdot 123} = \frac{\Delta_{54}}{\Delta_{55}} \\
x_{35\cdot 124} &= b_{35\cdot 124} = \frac{-\Delta_{53}}{\Delta_{55}} \\
x_{25\cdot 134} &= b_{25\cdot 134} = \frac{\Delta_{52}}{\Delta_{55}} \\
x_{15\cdot 234} &= b_{15\cdot 234} = \frac{-\Delta_{51}}{\Delta_{55}}.
\end{aligned} \tag{4}
$$

An examination of (4) shows that the notation for the solution is much more precise than the notation for the determinants. In addition, the notation for the solution is the more satisfactory inclusion notation; that of the determinants is an exclusion notation. An inclusion notation for determinants is indicated in the next chapter.

The formal use of Cramer's rule in the solution of (4.1.1) calls for the evaluation of $p + 1$ determinants of the pth order. The five fourth-order determinants Δ_{51}, Δ_{52}, Δ_{53}, Δ_{54}, and Δ_{55} must be calculated when $p = 4$. For most calculational purposes a back solution with the use of the identities given in Chapter 8 is preferable once the elimination process has led to some value of x_i.

EXERCISES

Evaluate the following determinants with the use of the methods described in this chapter. (You may first wish to refer to section 10.2.)

1. $\begin{vmatrix} 4 & 8 \\ -2 & 3 \end{vmatrix}$ = 28

2. $\begin{vmatrix} 2 & 1 & -2 \\ 1 & 1 & 1 \\ -1 & -2 & 3 \end{vmatrix}$ = 6

3. $\begin{vmatrix} 2 & 1 & -6 \\ 1 & 1 & 2 \\ -1 & -2 & 12 \end{vmatrix}$ = 12

4. $\begin{vmatrix} 2 & -1 & 1 & -1 \\ -1 & 2 & -1 & 1 \\ 1 & -1 & 2 & -1 \\ -1 & 1 & -1 & 2 \end{vmatrix}$ = 1

5. $\begin{vmatrix} 2 & -1 & 1 & 1 \\ -1 & 2 & -1 & 1 \\ 1 & -1 & 2 & 1 \\ -1 & 1 & -1 & 1 \end{vmatrix}$

6. $\begin{vmatrix} 1 & 1 & -1 \\ 1 & 2 & -3 \\ 2 & -1 & -2 \end{vmatrix}$

7. $\begin{vmatrix} 7 & 1 & -1 \\ 2 & 2 & -3 \\ 30 & -1 & -2 \end{vmatrix}$

8. $\begin{vmatrix} 2 & -1 & -2 \\ 0 & 1 & -1 \\ 1 & 1 & 1 \end{vmatrix}$

9. $\begin{vmatrix} 2 & -3 & -2 \\ 0 & 3 & -1 \\ 1 & -4 & 1 \end{vmatrix}$

10. $\begin{vmatrix} 1.0 & 0.4 & 0.5 & 0.6 \\ 0.4 & 1.0 & 0.3 & 0.4 \\ 0.5 & 0.3 & 1.0 & 0.2 \\ 0.6 & 0.4 & 0.3 & 1.0 \end{vmatrix}$

11. $\begin{vmatrix} 1.0 & 0.4 & 0.5 & 0.2 \\ 0.4 & 1.0 & 0.3 & 0.4 \\ 0.5 & 0.3 & 1.0 & 0.6 \\ 0.6 & 0.4 & 0.2 & 0.8 \end{vmatrix}$

12. $\begin{vmatrix} 1 & 1 & -1 \\ 2 & -1 & 2 \\ 3 & 0 & 1 \end{vmatrix}$

13. Show that the result of exercise 5 divided by that of exercise 4 is the value of x_4 in Table 4.2*b*.

14. Show that the result of exercise 7 divided by that of exercise 6 is the value of x_1 in Table 4.3*a*.

15. Show that the result of exercise 9 divided by that of exercise 8 is the negative of the value of y in Table 4.4*a*.

16. Show that the result of exercise 11 divided by that of exercise 10 is the ratio $\frac{0.36120}{0.30744}$ as indicated in Table 4.8*a*.

CHAPTER 10

The Evaluation of Determinants and Determinantal Ratios

10.1 Introduction. Though useful in determining the theoretical properties of the solutions of simultaneous equations, determinants are not very useful, at least with the conventional computational methods, in obtaining the numerical solutions needed in applied fields. This is in part due to the fact that many of the equation techniques usually presented are antiquated and too frequently trivial. We may know the usual evaluation procedures thoroughly and yet be unprepared to undertake such problems as we encounter in the subject of least squares. The classical solution of linear simultaneous equations by determinants calls for the evaluation of $p + 1$ determinants whereas elimination methods demand only the direct simultaneous calculation of two determinants and the evaluation of the other terms by the back solution.

10.2 The classical method. The classical method of evaluation of determinants of order greater than 3 is to reduce many of the elements of some row (or column) to zero by the continued use of the properties outlined in section 9.3, and then to expand in terms of the elements of that row (or column) by the method of section 9.4.

The computational process features determinants of decreasing order. As an illustration we calculate the value of

$$\Delta = \begin{vmatrix} 2 & 1 & -3 & 4 \\ 3 & 2 & 2 & 1 \\ -2 & -1 & 1 & 3 \\ 4 & -3 & 2 & 1 \end{vmatrix}. \tag{1}$$

If we multiply the elements of the second column by -2 and add to those of the first column, by 3 and add to the elements of the third column, and by -4 and add to the elements of the fourth column, we get

$$\Delta = \begin{vmatrix} 0 & 1 & 0 & 0 \\ -1 & 2 & 8 & -7 \\ 0 & -1 & -2 & 7 \\ 10 & -3 & -7 & 13 \end{vmatrix}. \tag{2}$$

Expanding in terms of the elements of the first row, we have

$$\Delta = -\begin{vmatrix} -1 & 8 & -7 \\ 0 & -2 & 7 \\ 10 & -7 & 13 \end{vmatrix}. \tag{3}$$

We now multiply the first row by 10 and add to the third row, so that

$$\Delta = -\begin{vmatrix} -1 & 8 & -7 \\ 0 & -2 & 7 \\ 0 & 73 & -57 \end{vmatrix} \tag{4}$$

whence

$$\Delta = -(-1)\begin{vmatrix} -2 & 7 \\ 73 & -57 \end{vmatrix} = -397. \tag{5}$$

10.3 The method of multiplication and subtraction. One use of the method of multiplication and subtraction is in obtaining the values of the determinant. For instance, let us consider the general determinant of order 4:

$$\Delta = \begin{vmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{vmatrix}. \tag{1}$$

We multiply the elements of the second row, those of the third row, and those of the fourth row in turn by a_{11} and compensate by multiplying the determinant by $1/a_{11}$. We then multiply the elements of the first row by $-a_{21}$ and add to the elements of the second row, by $-a_{31}$ and add to the elements of the third row, and by $-a_{41}$ and add to the elements of the fourth row, with the result

$$\Delta = \frac{1}{a_{11}^3}\begin{vmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ 0 & m_{22\cdot 1} & m_{23\cdot 1} & m_{24\cdot 1} \\ 0 & m_{32\cdot 1} & m_{33\cdot 1} & m_{34\cdot 1} \\ 0 & m_{42\cdot 1} & m_{43\cdot 1} & m_{44\cdot 1} \end{vmatrix}. \tag{2}$$

Expanding (2), we get

$$\Delta = \frac{1}{a_{11}^2}\begin{vmatrix} m_{22\cdot 1} & m_{23\cdot 1} & m_{24\cdot 1} \\ m_{32\cdot 1} & m_{33\cdot 1} & m_{34\cdot 1} \\ m_{42\cdot 1} & m_{43\cdot 1} & m_{44\cdot 1} \end{vmatrix}. \tag{3}$$

We treat the determinant of (3) in a similar fashion and arrive at

$$\Delta = \frac{1}{a_{11}^{\,2} m_{22\cdot 1}} \begin{vmatrix} m_{33\cdot 12} & m_{34\cdot 12} \\ m_{43\cdot 12} & m_{44\cdot 12} \end{vmatrix} \tag{4}$$

and therefore

$$\Delta = \frac{m_{44\cdot 123}}{a_{11}^{\,2} m_{22\cdot 1}}. \tag{5}$$

Continued application of the same reasoning shows that

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1h} & a_{1j} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2h} & a_{2j} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3h} & a_{3j} \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ a_{h1} & a_{h2} & a_{h3} & \cdots & a_{hh} & a_{hj} \\ a_{i1} & a_{i2} & a_{i3} & \cdots & a_{ih} & a_{ij} \end{vmatrix} = \frac{m_{ij\cdot(h)}}{m_{11}^{\,h-1} m_{22\cdot 1}^{\,h-2} m_{33\cdot 12}^{\,h-3} \cdots m_{h-1,h-1\cdot(h-2)}}. \tag{6}$$

Formula (6) shows that the value of the determinant can be calculated by dividing $m_{ij\cdot(h)}$ by the products of certain powers of diagonal terms in the elimination equations. As an illustration this method is applied to (10.2.1). The row sum check is omitted for simplicity of presentation, but it is recommended for computational use.

2	1	−3	4
3	2	2	1
−2	−1	1	3
4	−3	2	1
	1	13	−10
	0	−4	14
	−10	16	−14
		−4	14
		146	−114
			−1588

$$\Delta = \frac{-1588}{2^2 1} = -397.$$

Abbreviated forms of the method of multiplication and subtraction may be used if desired. The fact that the numerator is exactly divisible by the denominator provides a useful checking feature.

The determinants of the coefficients of the various sets of equations of Chapter 4 can be obtained by this method. Thus the determinant of the coefficients of the equations of Table 4.2*a* is $\frac{60}{2^2 3} = 5$, that of the equations of Table 4.8 is $\frac{0.30744}{(1.0000)^2(0.8400)} = 0.3660$, etc.

Non-diagonal pivots may be used, but then we must consider the algebraic sign of the pivot, which is plus or minus according to the evenness or oddness of the sum of the row or column. Thus the determinant of the coefficients in Table 4.7*b* is $\frac{-5}{-(-1)^2[-(-1)]} = 5$.

10.4 The method of multiplication and subtraction with (exact) division. An improvement over the method of multiplication and subtraction in evaluating determinants with desk machines is the method of multiplication and subtraction with (exact) division. We first note that

$$(1)\quad \begin{vmatrix} a_{11} & a_{12} & a_{13} & a_{1j} \\ a_{21} & a_{22} & a_{23} & a_{2j} \\ a_{31} & a_{32} & a_{33} & a_{3j} \\ a_{i1} & a_{i2} & a_{i3} & a_{ij} \end{vmatrix}$$

$$= \frac{1}{a_{11}^3} \begin{vmatrix} a_{11} & a_{12} & a_{13} & a_{1j} \\ 0 & d_{22\cdot 1} & d_{23\cdot 1} & d_{2j\cdot 1} \\ 0 & d_{32\cdot 1} & d_{33\cdot 1} & d_{3j\cdot 1} \\ 0 & d_{i2\cdot 1} & d_{i3\cdot 1} & d_{ij\cdot 1} \end{vmatrix}$$

$$= \frac{1}{a_{11}^2} \begin{vmatrix} d_{22\cdot 1} & d_{23\cdot 1} & d_{2j\cdot 1} \\ d_{32\cdot 1} & d_{33\cdot 1} & d_{3j\cdot 1} \\ d_{i2\cdot 1} & d_{i3\cdot 1} & d_{ij\cdot 1} \end{vmatrix}$$

$$= \frac{1}{a_{11}^2 d_{22\cdot 1}} \begin{vmatrix} d_{22\cdot 1} d_{33\cdot 1} - d_{32\cdot 1} d_{23\cdot 1} & d_{22\cdot 1} d_{3j\cdot 1} - d_{32\cdot 1} d_{2j\cdot 1} \\ d_{22\cdot 1} d_{i3\cdot 1} - d_{i2\cdot 1} d_{23\cdot 1} & d_{22\cdot 1} d_{ij\cdot 1} - d_{i2\cdot 1} d_{2j\cdot 1} \end{vmatrix}$$

$$= \frac{1}{d_{22\cdot 1}} \begin{vmatrix} d_{33\cdot 12} & d_{3j\cdot 12} \\ d_{i3\cdot 12} & d_{ij\cdot 12} \end{vmatrix} = d_{ij\cdot 123}.$$

In the same way we can show in general that the determinant of (10.3.6) is

$$(2)\ \Delta = |a_{ij}| = \frac{1}{a_{11}^{h-1}} |d_{ij\cdot 1}| = \frac{1}{d_{22\cdot 1}^{h-2}} |d_{ij\cdot 12}| = \frac{1}{d_{33\cdot 12}^{h-3}} |d_{ij\cdot 123}|$$

$$= \cdots = \frac{1}{d_{h-1,h-1\cdot(h-2)}} |d_{ij\cdot(h-1)}| = d_{ij\cdot(h)}.$$

This result indicates:

(*a*) That the determinant Δ can be evaluated directly by this method.
(*b*) That every d can be interpreted as a determinant.
(*c*) That all the computational entries of the method of multiplication and subtraction with (exact) division of Chapter 5 are themselves determinants.
(*d*) That all the rows and columns composing the determinant $d_{ij\cdot(h)}$ can be indicated by the subscripts. This notation, which is an inclusion notation, is much to be preferred to the usual exclusion notation since it specifies the elements of the determinant.

The illustration of the two previous sections is here solved with this method:

2	1	−3	4
3	2	2	1
−2	−1	1	3
4	−3	2	1
	1	13	−10
	0	−4	14
	−10	16	−14
		−2	7
		73	−57
			−397

This solution may be put in the compact form as in Table 5.9*b*. The problem then appears as

2	1	−3	4
3	2	2	1
−2	−1	1	3
4	−3	2	1
2	1	−3	4
3	1	13	10
−2	0	−2	7
4	−10	73	−397

Non-diagonal pivots may be used, but care must be taken to have the proper sign for each non-diagonal pivot. The solution of the illustrative problem is given when the pivots are chosen to coincide with those of section 10.2.

2	[1]	−3	4
3	2	2	1
−2	−1	1	3
4	−3	2	1
[−1]		8	−7
0		−2	7
10		−7	13
		[−2]	7
		73	−57
			−397

The order of elimination is not specified in the definition of a determinant so that the result does not depend on the order of elimination. Thus $d_{33\cdot 12} = d_{22\cdot 13} = d_{\cdot 123}$. Results of this type are demonstrated in section 5.7.

A more general notation is needed for general use with non-diagonal pivots. Aitken, in presenting the theory for non-diagonal pivots [A.1], suggests the umbral notation of Sylvester as the basis of a suitable notation. In this notation two rows are used, with the first row indicating the included rows of the matrix and the second indicating the included columns. Thus $d_{12\cdot 34} = \binom{134}{234}$ and $d_{33\cdot 12} = \binom{123}{123}$. The notation used thus far in this book is not general enough to indicate the determinant $\binom{134}{256}$. A logical extension of our notation indicates this by $d_{12,\,35,\,46}$, where the odd-numbered primary subscripts indicate the rows included and the even-numbered subscripts indicate the included columns. A secondary subscript indicates a row and the corresponding column. Thus

$$d_{34\cdot 1} = \binom{13}{14} \quad \text{and} \quad d_{23,\,45\cdot 1} = \binom{124}{135}.$$

Every d in Chapter 5 is the value of a determinant. Thus the value of the determinant of the coefficients of Table 5.8*b* is $d_{44\cdot 123} = 0.3660$.

More generally, each formula of Chapter 5 may be interpreted as a formula involving determinants. For example, (5.2.6) may be written in terms of the exclusion notation as

$$\Delta\Delta_{ij\cdot hk} = \Delta_{hh}\Delta_{ij} - \Delta_{hi}\Delta_{jh}$$

which is a special case of a frequently used theorem on determinants [A.2].

The value of $m_{ij\cdot(h)}$ in terms of determinants is given in (5.4.4) whereas (5.4.3) expresses the value of the determinant $d_{ij\cdot(h)}$ in terms of m's. Because the inclusion notation connected with the d's is to be preferred in many applications to the less precise conventional exclusion notation, the d's are used in the remainder of this book to indicate determinants.

Any method of solution of simultaneous equations (and evaluation of determinants) that yields computational entries that are always determinants may be called a *method of determinants*. The method of multiplication and subtraction with (exact) division is hence a method of determinants. Since none of the other computational methods of this book is a method of determinants and since this method seems superior to other methods of determinants, it is appropriate to refer to the method of multiplication and subtraction with (exact) division as *the method of determinants*.

10.5 A non-pivotal method of determinants. One early proponent of a method of multiplication and subtraction with (exact) division was Dodgson, who described his method in 1866 [B]. However, he used different elements, rather than a fixed pivot, as the multiplier and divisor in his elimination procedure. His method may be described as a condensation method and a method of determinants, but a non-pivotal one. A brief description of Dodgson's non-pivotal method is presented to show its similarity to the pivotal method described in the preceding section, although the pivotal method of determinants seems to lend itself to greater ease of computation.

Dodgson used a series of $ab - cd$ operations in which the elements a, b, c, and d are taken from adjacent rows and columns. Thus to evaluate the determinant

$$(1) \qquad \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} \qquad a_{22} \neq 0$$

we compute and record the values of $a_{11}a_{22} - a_{21}a_{12}$, $a_{12}a_{23} - a_{22}a_{13}$, $a_{21}a_{32} - a_{31}a_{22}$, $a_{22}a_{33} - a_{32}a_{23}$. It is to be noted that a_{22} is the only

element common to the four terms. It can be shown that the determinant (1) is

$$(2) \qquad \frac{1}{a_{22}} \begin{vmatrix} a_{11}a_{22} - a_{21}a_{12} & a_{12}a_{23} - a_{22}a_{13} \\ a_{21}a_{32} - a_{31}a_{22} & a_{22}a_{33} - a_{32}a_{23} \end{vmatrix}.$$

An illustration is given in (2). It is desired to find the determinant of the first three rows.

(2)

2	1	−3
3	2	2
−2	−1	1
	1	8
	1	4
		−2

Determinants of higher order are reduced by continued applications of essentially this same technique. The evaluation of the determinant in sections 10.2 and 10.3 is presented as an example.

2	1	−3	4
3	2	2	1
−2	−1	1	3
4	−3	2	1
	1	8	−11
	1	4	5
	10	1	−5
		−2	42
		39	−25
			−397

10.6 The method of single division. The methods so far presented are exact methods. In some cases approximate methods are preferable. A basic formula is first derived for evaluating determinants by the method of single division or by one of its variations. In the next section it is shown how the square root method can be used similarly. Determinantal formulas can be derived for the other methods described in this and the following section.

Consider the determinant of (10.4.1). We divide the first row by its leading element and get

$$(1) \qquad d_{ij\cdot 123} = a_{11}\begin{vmatrix} 1 & b_{12} & b_{13} & b_{1j} \\ a_{21} & a_{22} & a_{23} & a_{2j} \\ a_{31} & a_{32} & a_{33} & a_{3j} \\ a_{i1} & a_{i2} & a_{i3} & a_{ij} \end{vmatrix}.$$

We then multiply the first row by $-a_{21}$ and add to the second row, by $-a_{31}$ and add to the third row, by $-a_{41}$ and add to the fourth row, with the result

$$(2) \qquad d_{ij\cdot 123} = a_{11}\begin{vmatrix} 1 & b_{12} & b_{13} & b_{1j} \\ 0 & g_{22\cdot 1} & g_{23\cdot 1} & g_{2j\cdot 1} \\ 0 & g_{32\cdot 1} & g_{33\cdot 1} & g_{3j\cdot 1} \\ 0 & g_{i2\cdot 1} & g_{i3\cdot 1} & g_{ij\cdot 1} \end{vmatrix}.$$

Therefore,

$$(3) \qquad d_{ij\cdot 123} = g_{11}\begin{vmatrix} g_{22\cdot 1} & g_{23\cdot 1} & g_{2j\cdot 1} \\ g_{32\cdot 1} & g_{33\cdot 1} & g_{3j\cdot 1} \\ g_{i2\cdot 1} & g_{i3\cdot 1} & g_{ij\cdot 1} \end{vmatrix}.$$

Continuing in the same fashion, we get

$$(4) \qquad d_{ij\cdot 123} = g_{11}g_{22\cdot 1}\begin{vmatrix} g_{33\cdot 12} & g_{3j\cdot 12} \\ g_{i3\cdot 12} & g_{ij\cdot 12} \end{vmatrix}$$

and

$$(5) \qquad d_{ij\cdot 123} = g_{11}g_{22\cdot 1}g_{33\cdot 12}g_{ij\cdot 123}.$$

By similar reasoning the value of the determinant of (10.3.6) is

$$(6) \qquad d_{ij\cdot(h)} = g_{11}g_{22\cdot 1}g_{33\cdot 12}\cdots g_{hh\cdot(h-1)}g_{ij\cdot(h)}.$$

Formula (6) is the computational formula that is to be used with any variation of the method of single division. [This formula is identical with (7.3.1). The proof given in Chapter 7 together with the identification of $d_{ij\cdot(h)}$ as the desired determinant by (10.4.2) may be used in place of the proof of (6) given above.]

The formula (10.6.6) is well known by those who solve least squares problems by Gauss-Doolittle methods. See [C].

The various illustrations of different forms of the method of single division of Chapter 6 may be examined now and various determinants evaluated by (6). Pivotal terms only are needed. For example, the

determinant of the coefficients of the equations of Table 4.8*a* is

$$(1.0000)(0.8400)(0.7381)(0.5903) = 0.366$$

as shown by Tables 6.4*a*, 6.4*b*, 6.4*c*, and 6.4*f*.

10.7 The square root method. Determinants can also be evaluated by the square root method. Formula (6) becomes, with the use of (6.5.1) and (6.5.2)

$$(1) \qquad d_{ij\cdot(h)} = s_{11}{}^2 s_{22\cdot 1}{}^2 s_{33\cdot 12}{}^2 \cdots s_{hh\cdot(h-1)}{}^2 s_{jj\cdot(h)} s_{ij\cdot(h)}.$$

If $i = j = h + 1$, we have the formula

$$(2) \qquad d_{h+1,\,h+1\cdot(h)} = (s_{11} s_{22\cdot 1} s_{33\cdot 12} \cdots s_{hh\cdot(h-1)} s_{h+1,\,h+1\cdot(h)})^2.$$

Thus from Table 6.5*b* we see that

$$d_{44\cdot 123} = [(1.0000)(0.9165)(0.8591)(0.7683)]^2 = 0.3660.$$

10.8 The evaluation of partially symmetric determinants by symmetric methods. We may define a *partially symmetric determinant* to be one in which there is a symmetric minor, of order at least two, of at least one element of the principal diagonal. If this minor has an order one less than the order of the determinant, we may call the determinant *almost symmetric*. Thus the determinant

$$\begin{vmatrix} a_{11} & a_{12} & a_{1j} \\ a_{21} & a_{22} & a_{2j} \\ a_{i1} & a_{i2} & a_{ij} \end{vmatrix}$$

is almost symmetric if $a_{21} = a_{12}$, since the minor of a_{ij} is then symmetric.

We wish to make use of symmetric methods in the evaluation of partially symmetric determinants. This is done by inserting columns, for computational use, to make the matrix of the first p rows and p columns symmetric. Thus we might evaluate $d_{3j\cdot 12}$ with $a_{21} = a_{12}$, but with $a_{i1} \neq a_{1j}$, $a_{i2} \neq a_{2j}$, and with $a_{13} = a_{i1}$ and $a_{23} = a_{i2}$ by the computational form

$$\begin{array}{cccc} a_{11} & a_{12} & a_{13} & a_{ij} \\ * & a_{22} & a_{23} & a_{2j} \\ * & * & \cdots & a_{3j}. \end{array}$$

The application of the method of determinants (symmetric) gives

$$\begin{array}{cccc} d_{22\cdot 1} & d_{23\cdot 1} & \cdots & d_{2j\cdot 1} \\ & & \cdots & d_{3j\cdot 1} \\ \hline & & & d_{3j\cdot 12} \end{array}$$

whereas the Gauss-Doolittle method gives

$$\begin{array}{cccc} a_{11} & a_{12} & a_{13} & a_{1j} \\ 1 & b_{12} & b_{13} & b_{1j} \\ \hline & g_{22\cdot 1} & g_{23\cdot 1} & g_{2j\cdot 1} \\ & 1 & b_{23\cdot 1} & b_{2j\cdot 1} \\ \hline & & \cdots & g_{3j\cdot 12} \end{array}$$

so that

$$d_{3j\cdot 12} = g_{11}g_{22\cdot 1}g_{3j\cdot 12}.$$

If symmetry were lacking in two rows, it would be necessary to introduce two computational columns, etc.

We sometimes desire to evaluate a number of determinants that are alike except for the last column. If the number of these determinants is small it is advisable to use the foregoing method, with an additional column for each determinant. The evaluation of

$$d_{ij\cdot 123} = \begin{vmatrix} 1.0 & 0.4 & 0.5 & a_{1j} \\ 0.4 & 1.0 & 0.3 & a_{2j} \\ 0.5 & 0.3 & 1.0 & a_{3j} \\ 0.6 & 0.4 & 0.2 & a_{4j} \end{vmatrix}$$

with

$$\begin{array}{llll} a_{1j} = 0.6, & a_{2j} = 0.4, & a_{3j} = 0.2, & a_{4j} = 1.0 \\ a_{1j} = 0.2, & a_{2j} = 0.4, & a_{3j} = 0.6, & a_{4j} = 0.8 \\ a_{1j} = 0.8, & a_{2j} = 0.6, & a_{3j} = 0.4, & a_{4j} = 0.2 \end{array}$$

with the compact method of determinants (symmetric) is given in Table 10.8*a*.

If many of these determinants are to be evaluated, use a column for the coefficients a_{1j}, another for those of a_{2j}, another for those of a_{3j}, etc., and substitute the different values of a_{ij} in the results. The values of these coefficients are calculated with the method of determinants in Table 10.8*b*.

It follows that the value of $d_{4j\cdot 123}$ is

$$d_{4j\cdot 123} = -0.3700a_{1j} - 0.1300a_{2j} + 0.1000a_{3j} + 0.6200a_{4j}.$$

TABLE 10.8*a*

EVALUATION OF ALMOST SYMMETRIC DETERMINANTS

						Sum
1.0	0.4	0.5	0.6	0.2	0.8	3.5
*	1.0	0.3	0.4	0.4	0.6	3.1
*	*	1.0	0.2	0.6	0.4	3.0
*	*	*	1.0	0.8	0.2	3.2
1.0	0.4	0.5	0.6	0.2	0.8	3.5
	0.84	0.10	0.16	0.32	0.28	1.70
		0.620	−0.100	0.388	−0.028	0.880
			0.3660	0.4300	−0.2100	0.5860

TABLE 10.8*b*

SIMULTANEOUS EVALUATION OF ALMOST SYMMETRIC DETERMINANTS

				a_{1j}	a_{2j}	a_{3j}	a_{4j}	Sum
1.0	0.4	0.5	0.6	1	0	0	0	3.5
*	1.0	0.3	0.4	0	1	0	0	3.1
*	*	1.0	0.2	0	0	1	0	3.0
*	*	*	1.0	0	0	0	1	3.2
1.0	0.4	0.5	0.6	1	0	0	0	3.5
	0.84	0.10	0.16	−0.40	1.00	0	0	1.70
		0.620	−0.100	−0.380	−0.100	0.840	0	0.880
			0.3660	−0.3700	−0.1300	0.1000	0.6200	0.5860

Thus $d_{4j\cdot 123} = 0.4300$ when $a_{1j} = 0.2$, $a_{2j} = 0.4$, $a_{3j} = 0.6$, $a_{4j} = 0.8$. This is really a scheme for calculating the cofactors of the elements of the last column of (4).

10.9 The evaluation of determinantal ratios. The ratio of two determinants, rather than the value of a given determinant, is desired in solving simultaneous equations. The methods outlined in this chapter are useful in solving determinantal ratios, and indeed their application leads directly to the methods of solution of Chapters 4 to 6. For example, the evaluation of $b_{45\cdot 123}$ can be obtained from the determinantal ratio

(1) $$b_{45\cdot 123} = \frac{d_{45\cdot 123}}{d_{44\cdot 123}},$$

and the interpretation of the d's as determinants gives a statement of Cramer's rule in a precise notation. The expansion of the d's in terms of the g's and s's gives

(2) $$b_{45\cdot 123} = \frac{g_{11}g_{22\cdot 1}g_{33\cdot 12}g_{45\cdot 123}}{g_{11}g_{22\cdot 1}g_{33\cdot 12}g_{44\cdot 123}} = \frac{g_{45\cdot 123}}{g_{44\cdot 123}} = \frac{s_{45\cdot 123}}{s_{44\cdot 123}}.$$

Sometimes we can obtain simple formulas for determinantal ratios when the determinants are not of the same order. In multiple correlation theory, for example, we wish to evaluate the ratio of a determinant to one of its principal minors [D]. Formula (3) gives such a ratio in terms of the d's, m's, g's, and s's.

(3) $$\frac{d_{p,p\cdot(p-1)}}{d_{p-1,p-1\cdot(p-2)}} = \frac{m_{pp\cdot(p-1)}}{m_{11}m_{22\cdot 1}\cdots m_{p-1,p-1\cdot(p-2)}}$$

$$= g_{pp\cdot(p-1)} = s^2_{pp\cdot(p-1)}.$$

Table 6.4a shows, for example, $d_{44\cdot 123}/d_{33\cdot 12} = g_{44\cdot 123} = s^2_{44\cdot 123} = 0.5903$. This is an approximate result, but Table 5.10 shows that the exact value is 0.3660/0.6200.

The reader is referred to Aitken [E] for a further discussion of determinantal ratios. The amount to be added to the solution $b_{ij\cdot e}$, where e indicates a set of eliminated variables, as a result of the addition of a new variable x_h is a difference of determinantal ratios. Thus

(4) $$b_{ij\cdot eh} - b_{ij\cdot e} = \frac{d_{ij\cdot eh}}{d_{ii\cdot eh}} - \frac{d_{ij\cdot e}}{d_{ii\cdot e}}.$$

We expand $d_{ij\cdot eh}$ and $d_{ii\cdot eh}$ by definition, simplify the right-hand side to get

(5) $$b_{ij\cdot eh} - b_{ij\cdot e} = -\frac{d_{ih\cdot e}}{d_{ii\cdot e}}\frac{d_{hj\cdot ei}}{d_{hh\cdot ei}} = -b_{ih\cdot e}b_{hj\cdot ei}$$

and we have formal proof of (8.3.2).

10.10 The determination of all the principal minors of a determinant. We sometimes desire to obtain all the principal minors of a determinant. For example, we wish to write the characteristic equation of the matrix. It is not the purpose here to discuss in detail all the various methods available for finding the characteristic equation. The method of single division has been used by Reiersøl for writing the

principal minors [F]. The method emphasized here utilizes the method of determinants.

We start with a_{11} as a pivot and write all the $d_{ij\cdot 1}$ terms, the $d_{ij\cdot 2}$ terms, etc. After each diagonal element has been used as a pivot, we write the $d_{ij\cdot 12}$ terms, $d_{ij\cdot 13}$ terms, $d_{ij\cdot 23}$ terms, etc. The process is illustrated in Table 10.10*a*, where the principal minors are underscored. The first-order principal minors are the diagonal terms of the first block; the second-order principal minors are the diagonal terms of the three blocks below, etc.

TABLE 10.10*a*

THE PRINCIPAL MINORS OF A DETERMINANT

$\underline{a_{11}}$	a_{12}	a_{13}	a_{14}
a_{21}	$\underline{a_{22}}$	a_{23}	a_{24}
a_{31}	a_{32}	$\underline{a_{33}}$	a_{34}
a_{41}	a_{42}	a_{43}	$\underline{a_{44}}$
(a_{11})	$\underline{d_{22\cdot 1}}$	$d_{23\cdot 1}$	$d_{24\cdot 1}$
	$d_{32\cdot 1}$	$\underline{d_{33\cdot 1}}$	$d_{34\cdot 1}$
	$d_{42\cdot 1}$	$d_{43\cdot 1}$	$\underline{d_{44\cdot 1}}$
(a_{22})		$\underline{d_{33\cdot 2}}$	$d_{34\cdot 2}$
		$d_{43\cdot 2}$	$\underline{d_{44\cdot 2}}$
(a_{33})			$\underline{d_{44\cdot 3}}$
$(d_{22\cdot 1})$		$\underline{d_{33\cdot 12}}$	$d_{34\cdot 12}$
		$d_{43\cdot 12}$	$\underline{d_{44\cdot 12}}$
$(d_{33\cdot 1})$			$\underline{d_{44\cdot 13}}$
$(d_{33\cdot 2})$			$\underline{d_{44\cdot 23}}$
$(d_{33\cdot 12})$			$\underline{d_{44\cdot 123}}$

The different pivots are inserted at the left of each computation. A numerical illustration is given in Table 10.10*b*. The general plan is applicable to determinants of higher order.

TABLE 10.10*b*

THE PRINCIPAL MINORS OF A DETERMINANT—ILLUSTRATION

	0.570	0.210	−0.169	0.042
	0.184	0.195	−0.182	0.010
	−0.125	−0.265	0.522	−0.106
	−0.047	−0.016	−0.021	0.020
(0.570)		0.072510	−0.072644	−0.002028
		−0.124800	0.276415	−0.055170
		0.000750	−0.019913	0.013374
(0.195)			0.053560	−0.018020
			−0.007007	0.004060
(0.522)				0.008214
(0.072510)			0.019258	−0.007462
			−0.002438	0.001704
(0.276415)				0.004558
(0.053560)				0.000468
(0.019258)				0.000202

10.11 Determinants with complex elements. The method of determinants is also applicable, with modification, to determinants whose elements are complex numbers. If the element in the pth row and the qth column is indicated by $a'_{pq} + ia''_{pq}$, then the second-order determinant, $d_{pq\cdot 1}$, is

$$(1) \qquad d_{pq\cdot 1} = \begin{vmatrix} a'_{11} + ia''_{11} & a'_{1q} + ia''_{1q} \\ a'_{p1} + ia''_{p1} & a'_{pq} + ia''_{pq} \end{vmatrix}$$

$$= \begin{vmatrix} a'_{11} & a'_{1q} \\ a'_{p1} & a'_{pq} \end{vmatrix} + i \begin{vmatrix} a'_{11} & a''_{1q} \\ a'_{p1} & a''_{pq} \end{vmatrix} + i \begin{vmatrix} a''_{11} & a'_{1q} \\ a''_{p1} & a'_{pq} \end{vmatrix}$$

$$+ i^2 \begin{vmatrix} a''_{11} & a''_{1q} \\ a''_{p1} & a''_{pq} \end{vmatrix}.$$

It follows that $d_{pq\cdot 1}$ is of the form $d'_{pq\cdot 1} + id''_{pq\cdot 1}$, where

$$d'_{pq\cdot 1} = \begin{vmatrix} a'_{11} & a'_{1q} \\ a'_{p1} & a'_{pq} \end{vmatrix} - \begin{vmatrix} a''_{11} & a''_{1q} \\ a''_{p1} & a''_{pq} \end{vmatrix} \tag{2}$$

and

$$d''_{pq\cdot 1} = \begin{vmatrix} a'_{11} & a''_{1q} \\ a'_{p1} & a''_{pq} \end{vmatrix} + \begin{vmatrix} a''_{11} & a'_{1q} \\ a''_{p1} & a'_{pq} \end{vmatrix}. \tag{3}$$

The values of (2) and (3) are computed by operations that are sums and differences of $ab - cd$ and are themselves operational units of type U_3.

The basic computational formula for the third-order determinant $d_{pq\cdot 12}$ is given by the formula

$$d_{pq\cdot 12} = \frac{\begin{vmatrix} d'_{22\cdot 1} + id''_{22\cdot 1} & d'_{2q\cdot 1} + id''_{2q\cdot 1} \\ d'_{p2\cdot 1} + id''_{p2\cdot 1} & d'_{pq\cdot 1} + id''_{pq\cdot 1} \end{vmatrix}}{a'_{11} + ia''_{11}}. \tag{4}$$

The numerator of (4), a second-order determinant, is calculated by the method described above. It is of form $m' + im''$ and must be divided by $a'_{11} + ia''_{11}$ to get $d_{pq\cdot 12}$. It follows from the law for dividing complex numbers that

$$d_{pq\cdot 12} = \frac{a'_{11}m' + a''_{11}m'' + ia(a'_{11}m'' - a''_{11}m')}{(a'_{11})^2 + (a''_{11})^2}. \tag{5}$$

Here the right sides of

$$d'_{pq\cdot 12} = \frac{a'_{11}m' + a''_{11}m''}{(a'_{11})^2 + (a''_{11})^2} \tag{6}$$

and

$$d''_{pq\cdot 12} = \frac{a'_{11}m'' - a''_{11}m'}{(a'_{11})^2 + (a''_{11})^2} \tag{7}$$

are exactly divisible.

The general computational scheme for evaluation of a third-order determinant with complex elements is shown in Table 10.11a. The value $(a'_{11})^2 + (a''_{11})^2$ is inserted for ease of computation, at the left of the row in which m' and m'' appear.

The problem

$$d_{33\cdot 12} = \begin{vmatrix} 3 + i & 1 + 2i & -1 + 3i \\ 1 + 2i & 3 + i & 0 + 2i \\ 2 + 0i & 1 - i & 3 - 2i \end{vmatrix} \tag{8}$$

is used as an illustration in Table 10.11*a*. The m' results in 121, with m equal to -53, whereas $(a'_{11})^2 + (a''_{11})^2 = 10$. It follows that

$$d'_{33\cdot 12} = \frac{3(121) + 1(-53)}{10} = 31$$

$$d''_{33\cdot 12} = \frac{3(-53) - 1(121)}{10} = -28$$

so that

$$d_{33\cdot 12} = 31 - 28i.$$

This computational technique can be extended to determinants of higher order. The m' and m'', which are needed to compute the d' and d'', for each element after the first elimination, should be recorded. The m's might well be distinguished from the d's by using pencils or inks of different colors.

TABLE 10.11*a*

EVALUATION OF THIRD-ORDER DETERMINANT WITH COMPLEX ELEMENTS WITH THE METHOD OF DETERMINANTS

a'_{11} a'_{21} a'_{31}	a''_{11} a''_{21} a''_{31}	a'_{12} a'_{22} a'_{32}	a''_{12} a''_{22} a''_{32}	a'_{13} a'_{23} a'_{33}	a''_{13} a''_{23} a''_{33}
		$d'_{22\cdot 1}$ $d'_{32\cdot 1}$	$d''_{22\cdot 1}$ $d''_{32\cdot 1}$	$d'_{23\cdot 1}$ $d'_{33\cdot 1}$	$d''_{23\cdot 1}$ $d''_{33\cdot 1}$
$(a'_{11})^2 + (a''_{11})^2$				$m'_{33\cdot 12}$ $d'_{33\cdot 12}$	$m''_{33\cdot 12}$ $d''_{33\cdot 12}$

ILLUSTRATION

3 1 2	1 2 0	1 3 1	2 1 -1	-1 0 3	3 2 -2
		11 2	2 -6	5 13	5 -9
10				121 31	-53 -28

Determinants of complex numbers can also be evaluated by other methods. The reader is referred to the paper by Crout [G], where the method of single division is applied to the solution of simultaneous equations with complex coefficients. Such solutions are of course determinantal ratios rather than determinants, but determinantal results are available with the use of (10.6.6), which is valid for complex elements.

10.12 Determinants with approximate numbers as elements. The usual formulas given and those presented thus far are inadequate for suitable application to problems in which the elements of the determinant are themselves approximate digital numbers, the situation in most applications. The results, at least in so far as first-order errors are concerned, are very simple and easily computed by using the adjoint or inverse matrix. For a more detailed presentation, the reader is referred to the paper by Etherington, "On Errors in Determinants" [H.1], which was published in 1932. A much earlier paper on the allied question of solution of simultaneous equations with coefficients subject to error [H.2] was published in 1913.

One method of approach is to consider each element of the determinant as a range number or an error number and to use the general methods of Chapter 2 in connection with the specific methods of evaluating determinants outlined in this chapter. As an illustration we consider the determinant

$$\Delta = \begin{vmatrix} 2 & 4 & 6 \\ 1 & 3 & 5 \\ 2 & 1 & 4 \end{vmatrix} \tag{1}$$

where each element is a significant number, that is, the error in each number may be as large as one-half unit in the last digital position. In the notation of range numbers a determinant of the third order appears in the form

$$\begin{bmatrix} \Delta_H \\ \Delta_L \end{bmatrix} = \begin{vmatrix} a_{11H} & a_{12H} & a_{13H} \\ a_{11L} & a_{12L} & a_{13L} \\ a_{21H} & a_{22H} & a_{23H} \\ a_{21L} & a_{22L} & a_{23L} \\ a_{31H} & a_{32H} & a_{33H} \\ a_{31L} & a_{32L} & a_{33L} \end{vmatrix} \tag{2}$$

so that the determinant (1) takes the form

$$(3) \qquad \begin{bmatrix} \Delta_H \\ \Delta_L \end{bmatrix} = \begin{vmatrix} 2.5 & 4.5 & 6.5 \\ 1.5 & 3.5 & 5.5 \\ \\ 1.5 & 3.5 & 5.5 \\ 0.5 & 2.5 & 4.5 \\ \\ 2.5 & 1.5 & 4.5 \\ 1.5 & 0.5 & 3.5 \end{vmatrix}.$$

If the approximation-error numbers are used, the general form is

$$(4) \qquad \Delta \pm \epsilon(\Delta) = \begin{vmatrix} a_{11} \pm \epsilon_{11} & a_{12} \pm \epsilon_{12} & a_{13} \pm \epsilon_{13} \\ a_{21} \pm \epsilon_{21} & a_{22} \pm \epsilon_{22} & a_{23} \pm \epsilon_{23} \\ a_{31} \pm \epsilon_{31} & a_{32} \pm \epsilon_{32} & a_{33} \pm \epsilon_{33} \end{vmatrix}$$

and the illustration becomes

$$(5) \qquad \Delta \pm \eta(\Delta) = \begin{vmatrix} 2.0(5) & 4.0(5) & 6.0(5) \\ 1.0(5) & 3.0(5) & 5.0(5) \\ 2.0(5) & 1.0(5) & 4.0(5) \end{vmatrix}.$$

The use of approximation-error numbers rather than range numbers is advised. We first note that the value of (1) with each element treated as an exact element is 8. We next evaluate (5) by the method of determinants. We note further by (2.8.1) and (2.8.5) that

$$(6) \qquad \eta(x_1x_2 - x_3x_4) \leq |x'_1|\eta_2 + |x'_2|\eta_1 + |x'_3|\eta_4 + |x'_4|\eta_3$$

so that, if $\eta_1 = \eta_2 = \eta_3 = \eta_4 = \eta$, we have

$$(7) \qquad \eta(ab - cd) \leq \{|a| + |b| + |c| + |d|\}\eta.$$

Now

$$\begin{aligned} d_{22\cdot 1} &= 2.0(5) \times 3.0(5) - 1.0(5) \times 4.0(5) \\ &= [(2.0)(3.0) - (1.0)(4.0)] \pm (2.0 + 3.0 + 1.0 + 4.0)(0.5) \\ &= 2.0(50). \end{aligned}$$

Similarly,

$$d_{23\cdot 1} = 4.0(70), \qquad d_{32\cdot 1} = -6.0(45), \qquad d_{33\cdot 1} = -4.0(70).$$

It follows that

$$d_{33\cdot 12} = \frac{[2.0(50)][-4.0(70)] - [-6.0(45)][4.0(70)]}{2.0(5)}.$$

The calculation of an upper bound for the numerator can be performed with the use of (6) and (2.8.9). The bound for the error in the numerator is equal to or less than

$$(2.0)(7.0) + (4.0)(5.0) + (6.0)(7.0) + (4.0)(4.5) = 94.0$$

and

$$\Delta = d_{33\cdot 12} = \frac{16.0(940)}{2.0(5)}.$$

By using (2.8.9) we get

$$d_{33\cdot 12} = 8.0(490). \tag{8}$$

The computational arrangement is given in Table 10.12*a*. The results based on first-order errors provide upper bounds rather than limits for the value of the error of the determinant since (6) and (2.8.9) are inequalities rather than equalities. Experience shows that different orders of elimination may yield different bounds. In a general problem the actual error of the determinant $\epsilon(\Delta)$ may be much less than that indicated by the bound $\eta(\Delta)$.

TABLE 10.12*a*

CALCULATION OF A THIRD-ORDER DETERMINANT WITH APPROXIMATE ELEMENTS (METHOD OF DETERMINANTS)

2.0(5)	4.0(5)	6.0(5)	
1.0(5)	3.0(5)	5.0(5)	
2.0(5)	1.0(5)	4.0(5)	
	2.0(50)	4.0(70)	
	−6.0(45)	−4.0(70)	
		$\frac{16.0(940)}{2.0(5)}$	= 8(49)

Another method, more satisfactory in some respects than the method just given, is the method of incomplete numbers, if supplemented with a separate calculation of the bounds for the error.

We consider a determinant of approximation-error numbers such as (4). It can be expanded to give the sum of the eight determinants:

$$(9)\quad \Delta = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} + \begin{vmatrix} \epsilon_{11} & a_{12} & a_{13} \\ \epsilon_{21} & a_{22} & a_{23} \\ \epsilon_{31} & a_{32} & a_{33} \end{vmatrix} + \begin{vmatrix} a_{11} & \epsilon_{12} & a_{13} \\ a_{12} & \epsilon_{22} & a_{23} \\ a_{13} & \epsilon_{23} & a_{33} \end{vmatrix}$$
$$+ \begin{vmatrix} a_{11} & a_{12} & \epsilon_{13} \\ a_{21} & a_{22} & \epsilon_{23} \\ a_{31} & a_{32} & \epsilon_{33} \end{vmatrix} + \begin{vmatrix} \epsilon_{11} & \epsilon_{12} & a_{13} \\ \epsilon_{21} & \epsilon_{22} & a_{23} \\ \epsilon_{31} & \epsilon_{32} & a_{33} \end{vmatrix} + \begin{vmatrix} \epsilon_{11} & a_{12} & \epsilon_{13} \\ \epsilon_{21} & a_{22} & \epsilon_{23} \\ \epsilon_{31} & a_{32} & \epsilon_{33} \end{vmatrix}$$
$$+ \begin{vmatrix} a_{11} & \epsilon_{12} & \epsilon_{13} \\ a_{21} & \epsilon_{22} & \epsilon_{23} \\ a_{31} & \epsilon_{32} & \epsilon_{33} \end{vmatrix} + \begin{vmatrix} \epsilon_{11} & \epsilon_{12} & \epsilon_{13} \\ \epsilon_{21} & \epsilon_{22} & \epsilon_{23} \\ \epsilon_{31} & \epsilon_{32} & \epsilon_{33} \end{vmatrix}.$$

The first of these determinants is the determinant of the incomplete numbers. The second, third, and fourth give terms that contribute to the first-order errors; the fifth, sixth, and seventh yield second-order error terms; the last gives third-order error terms. We limit the discussion to the first-order error terms although terms of higher order can be computed and should be if the errors are relatively large.

Formal expansion of the first-order terms shows that

$$(10)\qquad \epsilon(\Delta) = \sum_i \sum_j (-1)^{i+j}\, \epsilon_{ij}\, \Delta_{ij},$$

where $\epsilon(\Delta)$ is the first-order error term of the determinant and Δ_{ij} is the minor of a_{ij}. Moreover (10) holds not only for determinants of order three, but also for determinants of any order. It is not difficult then to find the first-order approximation to a determinant if the error of each element is known.

However, this is not the usual situation. The error of each element is usually not known, but is specified as having a maximum absolute value of η. In problems involving determinants of high order, it is not a trivial task to get the minor of every element.

We adapt (10) to the first problem by using the formula

$$(11)\qquad \eta(\Delta) \leq \sum_i \sum_j \eta_{ij} \left| \Delta_{ij} \right|$$

in accordance with the method of Chapter 2. The difficulty in connection with the second problem is eased appreciably by making use of methods of computing the inverse matrix described in Chapter 13.

If all $\eta_{ij} = \eta$, we have

$$(12)\qquad \eta(\Delta) = \eta \sum_i \sum_j \left| \Delta_{ij} \right|.$$

The minors of a third-order determinant are easily evaluated. We first note that the incomplete number evaluation of the determinant of (5) is 8. It is necessary only to get the values

$$\begin{matrix} \Delta_{11} & \Delta_{21} & \Delta_{31} \\ \Delta_{12} & \Delta_{22} & \Delta_{32} \\ \Delta_{13} & \Delta_{23} & \Delta_{33} \end{matrix}$$

to take the absolute value of each, to add and multiply by one-half unit to get a bound for the error. Now Δ_{22} is $d_{33\cdot 12}$, Δ_{32} is $d_{23\cdot 1}$, Δ_{23} is $d_{32\cdot 1}$, and Δ_{33} is $d_{22\cdot 1}$. We need in addition Δ_{11}, Δ_{21}, Δ_{31}, Δ_{12}, Δ_{13}. The calculations needed for the evaluation of the incomplete numbers and the minors for a third-order determinant are compactly arranged in the following scheme:

TABLE 10.12*b*

FORM FOR CALCULATION OF APPROXIMATE THIRD-ORDER DETERMINANTS

$a_{11}(\eta_{11})$	$a_{12}(\eta_{12})$	$a_{13}(\eta_{13})$
$a_{21}(\eta_{21})$	$a_{22}(\eta_{22})$	$a_{23}(\eta_{23})$
$a_{31}(\eta_{31})$	$a_{32}(\eta_{32})$	$a_{33}(\eta_{33})$
Δ_{11}	Δ_{21}	Δ_{31}
Δ_{12}	$d_{22\cdot 1}$	$d_{23\cdot 1}$
Δ_{13}	$d_{32\cdot 1}$	$d_{33\cdot 1}$
		$d_{33\cdot 12}$

The evaluation of (5) by this method is then

$2(\frac{1}{2})$	$4(\frac{1}{2})$	$6(\frac{1}{2})$
$1(\frac{1}{2})$	$3(\frac{1}{2})$	$5(\frac{1}{2})$
$2(\frac{1}{2})$	$1(\frac{1}{2})$	$4(\frac{1}{2})$
7	−6	−5
10	2	4
2	−6	−4
		8

so that the value of the determinant of the approximate numbers is

$$8 \pm \tfrac{1}{2}(7 + 6 + 5 + 10 + 2 + 4 + 2 + 6 + 4) = 8 \pm 23.$$

This method gives us closer bounds than the method of Table 10.12*a*.

If the numbers were accurate to one more decimal place, that is, $\eta_{ij} = 0.05$, the value of the determinant could be written as 8 ± 2.3; if it were accurate to two more decimal places, it would be 8 ± 0.23.

As a second illustration, consider the problem in which the a_{ij} are those of Table 10.12*a*, but the error terms are different, and may be different from each other. In this case apply (11). Thus

TABLE 10.12*c*

SECOND ILLUSTRATION

2.0(4)	4.0(2)	6.0(3)
1.0(1)	3.0(2)	5.0(2)
2.0(1)	1.0(2)	4.0(3)
7	−6	−5
10	2	4
2	−6	−4
		8

and a bound for the error is

$$7(0.4) + 6(0.2) + 5(0.3) + 10(0.1) + 2(0.2) + 4(0.2) + 2(0.1) + 6(0.2) + 4(0.3) = 10.3,$$

so that

$$\Delta = 8 \pm 10.3.$$

This method is easy to apply to determinant of the third order and for situations for which the minors of all elements are available. It is shown in Chapter 17 that the adjoint matrix or the inverse matrix may be used in finding the value of $|\Delta_{ij}|$.

REFERENCES

A. 1. A. C. Aitken, "On the evaluation of determinants, the formation of their adjugates, and the practical solution of simultaneous linear equations," *Proceedings Edinburgh Mathematical Society*, Series 2, **3**, 207–219 (1932).

2. M. Bocher, *Introduction to Higher Algebra*, The Macmillan Co., New York, 1907. See page 33.

B. C. L. Dodgson, "Condensation of determinants," *Proceedings of the Royal Society*, **15**, 150–155 (1866).

C. W. E. Deming, *The Statistical Adjustment of Data*, John Wiley and Sons, New York, 1943, p. 161.

D. P. S. Dwyer, "The evaluation of determinants," *Psychometrika*, **6**, 191–204 (1941). See pp. 203–204.

E. A. C. Aitken, *Determinants and Matrices*, Oliver and Boyd, London, 1942. See pp. 107–110.

F. O. Reiersøl, "A method for recurrent computation of all the principal minors of a determinant, and its application in confluence analysis," *Annals of Mathematical Statistics*, **11**, 193–198 (1940).

G. P. D. Crout, "A short method for evaluating determinants and solving systems of linear equations with real or complex coefficients," *Marchant Methods*, M.M. 182, September 1941, Marchant Calculating Machine Co., Oakland, Calif.

H. 1. I. M. H. Etherington, "On errors in determinants," *Proceedings of the Edinburgh Mathematical Society*, Series 2, **3**, 107–117 (1932).

2. F. R. Moulton, "On the solutions of linear equations having small determinants," *American Mathematical Monthly*, **20**, 242–249 (1913).

EXERCISES

1–12. Work exercises 1 to 12 of Chapter 9 by Dodgson's method.

13–24. Work exercises 1 to 12 of Chapter 9 by the method of determinants.

25–36. Work exercises 1 to 12 of Chapter 9 by some variant of the method of single division.

37. Evaluate the determinant of the coefficients of the equations of Table 4.8*a* by the square root method.

38. Solve the problem of Table 4.13*a* by the method of determinants.

39–43. Show how determinants can be used in solving exercises 5 to 9 of Chapter 9.

44. Use determinants in solving exercise 9.10.

45. Determine all the principal minors of the determinant of the coefficients of Table 4.13*a* by the method of section 10.10.

46. Determine all the principal minors of the determinant

$$\begin{vmatrix} 15 & 11 & 6 & -9 & -15 \\ 1 & 3 & 9 & -3 & -8 \\ 7 & 6 & 6 & -3 & -11 \\ 7 & 7 & 5 & -3 & -11 \\ 17 & 12 & 5 & -10 & -16 \end{vmatrix}$$

by the method of section 10.10.

47. Evaluate $\begin{vmatrix} 2+i & 3-2i \\ 4+3i & -2+i \end{vmatrix}$.

48. Evaluate $\begin{vmatrix} 4-3i & 6-2i \\ 2+3i & 8+5i \end{vmatrix}$.

49. Evaluate $\begin{vmatrix} 2+i & 0+i & 1+0i \\ 3-2i & 2-3i & 7-3i \\ 4+6i & -1+2i & 8+4i \end{vmatrix}$.

50–57. Use (10.12.12) in obtaining bounds for the errors of the determinants of exercises 9.1, 9.2, 9.3, 9.6, 9.7, 9.8, 9.9, and 9.12 if values of the elements are significant to one decimal position, that is, $\eta \leq 0.05$.

58. Use the material of Chapter 2 and the results of exercises 51 and 52 in determining a bound for the error of x_3 in

$$2x_1 + x_2 - 2x_3 = -6$$

$$x_1 + x_2 + x_3 = 2$$

$$-x_1 - 2x_2 + 3x_3 = 12.$$

CHAPTER 11

The Evaluation of Linear Forms

11.1 Introduction. Frequently we need to find the numerical value of a linear form when the specific values of the variables are given implicitly by a set of linear equations. For example, we may wish to determine the value of the linear form [A]

$$a_{i1}x_1 + a_{i2}x_2 + a_{i3}x_3 + a_{i4}x_4 = a_{i5}, \qquad i > 4 \tag{1}$$

when the values x_1, x_2, x_3, and x_4 are defined by (4.1.2).

The purpose of this chapter is to show how problems of this type can be solved by using the findings of earlier chapters. Special emphasis is given to solutions that yield the value of the form without finding the explicit values of the variables in the form.

In many problems simultaneous evaluation of a number of linear forms is desired. For example, we may wish to evaluate (1) for a different value of i. Application is made in this chapter to a number of situations of this sort.

The reader familiar with matrices will recognize that the problems of this chapter may be treated by matrices. He may prefer to develop the techniques with matrix methods shown in later chapters and so may wish to do no more than glance through the pages of this chapter.

11.2 The basic theory. The basic problem is illustrated by the four variable illustrations above. We could solve for x_1, x_2, x_3, x_4 by the methods of earlier chapters and substitute in (1), but a preferable method is to obtain the value a_{i5} without getting the values of x_1, x_2, x_3, and x_4. This means that (4.1.2), with $d_{44\cdot123} \neq 0$, and (11.1.1) must be satisfied simultaneously. The condition for this is that

$$d_{i5\cdot1234} = 0. \tag{1}$$

It follows of course that $m_{i5\cdot1234}$, $g_{i5\cdot1234}$, and $s_{i5\cdot1234}$ must also be zero, with $m_{44\cdot123}$, $g_{44\cdot123}$, and $s_{44\cdot123} \neq 0$. The extension to the general problem is obvious.

11.3 Exact methods. Consider the determinant

$$(1) \qquad \begin{vmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} \\ a_{i1} & a_{i2} & a_{i3} & a_{i4} & a_{i5} \end{vmatrix} = d_{i5\cdot 1234}$$

and also a determinant $d^0_{i5\cdot 1234}$ that is identical with $d_{i5\cdot 1234}$, except that a_{i5} is replaced by zero. Expanding each determinant in terms of the last row, we have

$$(2) \qquad d_{i5\cdot 1234} = a_{i5}d_{44\cdot 123} + \text{"terms"}$$

$$(3) \qquad d^0_{i5\cdot 1234} = 0\cdot d_{44\cdot 123} + \text{"terms"}$$

where the "terms" are the same. Subtracting (3) from (2) and solving for a_{i5}, we get, since $d_{i5\cdot 1234} = 0$,

$$(4) \qquad a_{i5} = \frac{-d^0_{i5\cdot 1234}}{d_{44\cdot 123}}.$$

A similar statement holds for the more general case. Thus

$$(5) \qquad a_{i,\,p+1} = -\frac{d^0_{i,\,p+1\cdot(p)}}{d_{pp\cdot(p-1)}},$$

where p is the number of simultaneous equations that define the x's implicitly.

The method of determinants is used in Table 11.3a to calculate the values of

$$(6) \qquad 0.400x_1 + 0.600x_2 + 0.800x_3 = ?$$

with

$$0.800x_1 + 0.480x_2 + 0.360x_3 = 1.000$$

$$0.480x_1 + 0.800x_2 + 0.360x_3 = 0.000$$

$$0.360x_1 + 0.360x_2 + 0.860x_3 = 0.000.$$

The formula becomes

$$(7) \qquad a_{i,\,p+1} = -\frac{m^0_{i,\,p+1\cdot(p)}}{a_{11}m_{22\cdot 1}m_{33\cdot 12}\cdots m_{p-1,\,p-1\cdot(p-2)}}.$$

TABLE 11.3*a*

IMPLICIT EVALUATION OF LINEAR FORMS WITH THE METHOD OF DETERMINANTS—GENERAL

a_{11}	a_{12}	a_{13}	a_{14}
a_{21}	a_{22}	a_{23}	a_{24}
a_{31}	a_{32}	a_{33}	a_{34}
a_{41}	a_{42}	a_{43}	0
	$d_{22\cdot1}$	$d_{23\cdot1}$	$d_{24\cdot1}$
	$d_{32\cdot1}$	$d_{33\cdot1}$	$d_{34\cdot1}$
	$d_{42\cdot1}$	$d_{43\cdot1}$	$d^0_{44\cdot1}$
		$d_{33\cdot12}$	$d_{34\cdot12}$
		$d_{34\cdot12}$	$d^0_{44\cdot12}$
			$d^0_{44\cdot123}$
			F

ILLUSTRATION

0.80	0.48	0.36	1.00
0.48	0.80	0.36	0.00
0.36	0.36	0.86	0.00
0.40	0.60	0.80	0.00
	0.4096	0.1152	−0.4800
	0.1152	0.5584	−0.3600
	0.2880	0.4960	−0.4000
		0.269312	−0.115200
		0.212480	−0.032000
			0.03872000
			$\frac{0.03872000}{0.269312} = -0.144$

11.4 Approximate methods. Approximate methods are also applicable. Application of (10.6.6) to (11.3.5) yields

(1) $$a_{i,\,p+1} = -g^0{}_{i,\,p+1\cdot(p)},$$

where $g^0_{i,\,p+1}$ is the term obtained by replacing $a_{i,\,p+1}$ by zero. It follows that

$$a_{i,\,p+1} = -[0 - g_{i1}b_{i,\,p+1} - g_{i2\cdot 1}b_{2,\,p+1\cdot 1} - \cdots - g_{ip\cdot(p-1)}b_{p,\,p+1\cdot(p-1)}]$$

so that

$$(2) \quad a_{i,\,p+1} = g_{i1}b_{i,\,p+1} + g_{i2\cdot 1}b_{2,\,p+1\cdot 1} + \cdots g_{ip\cdot(p-1)}b_{p,\,p+1\cdot(p-1)}.$$

This formula is applicable to any variation of the method of single division. If the equations are symmetric, the primary subscripts of the a's may be interchanged so that the Gauss-Doolittle form exhibits the two factors to be multiplied in successive couplet rows. The solution of the illustration of the problem used in this chapter is given in Table 11.4a,

TABLE 11.4a

EVALUATION OF LINEAR FORM—VARIATION OF THE METHOD OF SINGLE DIVISION

Compact Method of Single Division				Gauss-Doolittle Method				
0.800	0.480	0.360	1.000	0.800	0.480	0.360	1.000	0.400
0.480	0.800	0.360	0.000	*	0.800	0.360	0.000	0.600
0.360	0.360	0.860	0.000	*	*	0.860	0.000	0.800
0.400	0.600	0.800	0					
				0.800	0.480	0.360	1.000	0.400
0.800	0.600	0.450	1.250	1.000	0.600	0.450	1.250	0.500
0.480	0.512	0.281	−1.172					
0.360	0.144	0.658	−0.427		0.512	0.144	−0.600	0.360
0.400	0.360	0.519	(−0.144)		1.000	0.281	−1.172	0.703
						0.658	−0.281	0.519
						1.000	−0.427	0.789
								−0.144

with the use of the compact method of single division and the Gauss-Doolittle forms. In either case the value of the linear form is

$$(0.400)(1.250) + (0.360)(-1.172) + (0.519)(-0.427) = -0.144.$$

11.5 An alternative to the back solution. The ideas outlined in this chapter may be used to provide a substitute for the back solution since

each x_i is a special case of a linear form. Thus x_1 is a linear form with $a_{i1} = 1$, $a_{i2} = 0$, $a_{i3} = 0$. It is at once possible to find x_1, x_2, x_3, and x_4 from the forward solution by the methods just outlined.

The Gauss-Doolittle presentation of the solution of the problem of Table 4.8*a* by this method is shown in Table 11.5*a*.

TABLE 11.5*a*

VALUE OF x_i—GAUSS-DOOLITTLE METHOD—NO BACK SOLUTION

1.0	0.4	0.5	0.6	0.2	1.0	0	0	0
*	1.0	0.3	0.4	0.4	0	1.0	0	0
*	*	1.0	0.2	0.6	0	0	1.0	0
*	*	*	1.0	0.8	0	0	0	1.0
1.0000	0.4000	0.5000	0.6000	0.2000	1.0000	0.0000	0.0000	0.0000
1.0000	0.4000	0.5000	0.6000	0.2000	1.0000	0.0000	0.0000	0.0000
	0.8400	0.1000	0.1600	0.3200	−0.4000	1.0000	0.0000	0.0000
	1.0000	0.1190	0.1905	0.3810	−0.4762	1.1905	1.0000	0.0000
		0.7381	−0.1190	0.4619	−0.4524	−0.1190	1.0000	0.0000
		1.0000	−0.1612	0.6258	−0.6129	−0.1612	1.3548	0.0000
			0.5903	0.6935	−0.5966	−0.2097	0.1612	1.0000
			1.0000	1.1748	−1.0107	−0.3552	0.2731	1.6941
−0.9364	0.0602	0.8152	1.1748					
x_1	x_2	x_3	x_4					

where $x_1 = (1.0000)(0.2000) + (-0.4000)(0.3810) + (-0.4524)(0.6258) + (-0.5966)(1.1748) = -0.9364$, etc.

The same material is presented in Table 11.5*b* by the compact method of single division. The general presentation is given on the left, and the numerical illustration on the right. Identification of terms on the left is made through the continued use of (8.3.4).

The b's in the last four rows are solutions of groups of related equations. Thus in column four we have $b_{14\cdot23}$, $b_{24\cdot13}$, and $b_{34\cdot12}$. The entries of the rows of Table 11.5*a* are similarly identifiable.

The methods of determinants can be similarly used. However, the relations are such that the general ideas of this section are more useful in connection with some variation of the method of single division.

Topics related to those of this chapter are discussed further in Chapter 13.

TABLE 11.5*b*

VALUES OF x_i—COMPACT METHOD OF SINGLE DIVISION—NO BACK SOLUTION

General					Illustration				
a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	1.0000	0.4000	0.5000	0.6000	0.2000
a_{21}	a_{22}	a_{23}	a_{24}	a_{25}	0.4000	1.0000	0.3000	0.4000	0.4000
a_{31}	a_{32}	a_{33}	a_{34}	a_{35}	0.5000	0.3000	1.0000	0.2000	0.6000
a_{41}	a_{42}	a_{43}	a_{44}	a_{45}	0.6000	0.4000	0.2000	1.0000	0.8000
1	0	0	0	0	1.0000	0	0	0	0
0	1	0	0	0	0	1.0000	0	0	0
0	0	1	0	0	0	0	1.0000	0	0
0	0	0	1	0	0	0	0	1.0000	0
a_{11}	b_{12}	b_{13}	b_{14}	b_{15}	1.0000	0.4000	0.5000	0.6000	0.2000
a_{21}	$a_{22\cdot1}$	$b_{23\cdot1}$	$b_{24\cdot1}$	$b_{25\cdot1}$	0.4000	0.8400	0.1190	0.1905	0.3810
a_{31}	$a_{32\cdot1}$	$a_{33\cdot12}$	$b_{34\cdot12}$	$b_{35\cdot12}$	0.5000	0.1000	0.7381	−0.1613	0.6258
a_{41}	$a_{42\cdot1}$	$a_{43\cdot12}$	$a_{44\cdot123}$	$b_{45\cdot123}$	0.6000	0.1600	−0.1190	0.5903	1.1748
1	$-b_{12}$	$-b_{13\cdot2}$	$-b_{14\cdot23}$	(x_1)	1.0000	−0.4000	−0.4524	−0.5966	(−0.9364)
0	1	$-b_{23\cdot1}$	$-b_{24\cdot13}$	(x_2)	0	1.0000	−0.1190	−0.2097	(0.0602)
0	0	1	$-b_{34\cdot12}$	(x_3)	0	0	1.0000	0.1612	(0.8152)
0	0	0	1	(x_4)	0	0	0	1.0000	(1.1748)
x_1	x_2	x_3	x_4		−0.9364	0.0602	0.8152	1.1748	

REFERENCE

A. P. S. Dwyer, "The evaluation of linear forms," *Psychometrika*, **6**, 355–365 (1941).

EXERCISES

1. Given the system of equations

$$3x + 7y - 4z = 11$$

$$x + y - 8z = 2$$

$$4x + 5y + 6z = 23,$$

find the value of $2x - 3y + 4z$, without finding the values of x, y, and z, by the method of determinants.

2. Find the values of x, y, and z by the method of Table 11.4*a* for exercise 1.

3. Solve for x_1 in exercise 4.3 by the method of Table 11.4*a*.

4. Solve for x_1, x_2, x_3 in exercise 4.3 by the method of Table 11.5*a*.

5. Solve exercise 6.11 with the general technique of this chapter and by the square root method.

CHAPTER 12

An Introduction to the Algebra of Matrices

12.1 Introduction. Matrix algebra is of great use in handling problems of the type discussed in the preceding chapters. This chapter is inserted to introduce the reader not familiar with matrices to the basic algebra. As in the chapter on determinants, only the basic concepts and the manipulative theorems are presented. Formal proofs may be found in appropriate tests.

The generalization of arithmetic leads to the quantities of algebra. A first generalization of algebra leads to complex numbers, where each component of the dual number is itself an algebraic quantity. The next generalization leads to vectors, where each vector is a row, or a column, of algebraic or arithmetic quantities. The next generalization leads to columns of rows of these numbers, or rows of columns of them. An entity of this sort is called a *matrix* and represents a double generalization of the numbers of algebra and arithmetic.

12.2 Definitions and notation. The m rows and n columns of elements a_{ij}, $i = 1, 2, \cdots m$, $j = 1, 2 \cdots n$ when placed in a rectangular array is called a matrix or, more formally, an m by n matrix. Thus the matrix a of elements a_{ij} is

$$(1) \qquad a = [a_{ij}] = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2j} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{i1} & a_{i2} & \cdots & a_{ij} & \cdots & a_{in} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mj} & & a_{mn} \end{bmatrix}.$$

Symbols such as double lines, brackets, and parentheses are sometimes used to indicate matrices.

The term matrix may be more generally used to indicate any arrange-

ment, not necessarily rectangular, of elements. In mathematics the term is usually restricted to the rectangular array described above.

When $m = n$, the square array is called a square matrix of order n. If $m = n = 1$, we have the matrix with the single element a_{11}. This is interpreted as the algebraic quantity a_{11}.

The algebraic or arithmetic quantities that serve as elements of the matrix are known as *scalars*. Thus a one-by-one matrix is a scalar.

A matrix is arithmetic if every element is an arithmetic number. If one or more of the elements of the matrix is composed of algebraic numbers, the matrix is algebraic. Thus the matrix $\begin{bmatrix} 1 & 2 & 3 \\ 2 & 1 & 2 \\ 1 & -1 & 0 \end{bmatrix}$ is arithmetic whereas the matrix $\begin{bmatrix} x & 1 & 3 \\ 2 & x & 2 \\ 1 & -1 & x \end{bmatrix}$ is algebraic.

If $n = 1$, we have a single-column matrix called a column vector. If $m = 1$, we have a single-row matrix called a row vector.

Two matrices a and b are said to be equal if every element in the one equals the corresponding element in the other, that is, if $a_{ij} = b_{ij}$. It is necessary of course that they have identical rows and columns. Two matrices a and b are identical if $a_{ij} \equiv b_{ij}$.

A matrix is a null or zero matrix if every one of its elements is zero. It is represented by 0.

A square matrix having unity for each diagonal element and having zero for every other element is known as an identity matrix. It is represented by I. The order of the matrix may be used as a subscript. Thus I_4 indicates an identity matrix of order four.

A diagonal matrix is a square matrix with all elements zero except those of the principal diagonal, at least one of which is different from zero. Thus the identity matrix is a special case of a diagonal matrix.

A matrix may be formed from a by interchanging rows and columns. Such a matrix is known as the transpose of a. It is denoted by a^T or a'. If a is an m by n matrix, a' is an n by m matrix with the element from the ith row and the jth column of a in the jth row and the ith column of a'. By definition the transpose of a' is itself a. Thus $(a')' = a$.

A symmetric matrix is a square matrix in which $a = a'$. The definition implies that $a_{ij} = a_{ji}$ for all i, j. Thus the matrix

$$\begin{bmatrix} 1.0 & 0.4 & 0.6 \\ 0.4 & 1.0 & 0.2 \\ 0.6 & 0.2 & 1.0 \end{bmatrix} \tag{2}$$

is symmetric.

The sum of two or more matrices with corresponding numbers of rows and columns is defined to be the matrix of the sum of all the corresponding elements. Thus if

$$(3)\quad a_{ij} + b_{ij} = c_{ij}, \qquad \text{then} \qquad c = [c_{ij}] = [a_{ij} + b_{ij}]$$
$$= [a_{ij}] + [b_{ij}] = a + b = b + a.$$

Matrices cannot be added if the numbers of rows and columns do not correspond.

Subtraction is a special case of addition, with the elements of the subtracted matrix being multiplied by -1 before addition.

$$(4)\quad a_{ij} - b_{ij} = c_{ij}, \qquad \text{then} \qquad c = [c_{ij}] = [a_{ij} + (-b_{ij})] = a - b.$$

12.3 Matrix multiplication. In the simplest type of matrix multiplication the matrix is multiplied by a scalar. If we let a represent the matrix and k the scalar, we define the product

$$(1)\qquad ka = ak = [ka_{ij}] = [a_{ij}k].$$

Thus

$$k\begin{bmatrix} -2 & 3 \\ 1 & 4 \end{bmatrix} = \begin{bmatrix} -2k & 3k \\ k & 4k \end{bmatrix}.$$

It follows that the product is 0 if either $k = 0$ or $a = 0$.

The more general type of matrix multiplication involves the product of two matrices. If a is an m by n matrix and b is the n by p matrix, the definition of the product is the m by p matrix,

$$(2)\qquad ab = \sum_{k=1}^{n} a_{ik}b_{kj}.$$

The element in the ith row and the jth column of the product is obtained by multiplying the elements in the ith row of a by those in the jth column of b and then adding the results. Thus

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix}\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \\ a_{31}b_{11} + a_{32}b_{21} & a_{31}b_{12} + a_{32}b_{22} \end{bmatrix}$$

whereas

$$\begin{bmatrix} 5 & 4 \\ 1 & 3 \end{bmatrix}\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 5 & 4 & 0 \\ 1 & 3 & 0 \end{bmatrix}$$

and

$$\begin{bmatrix} 5 & 4 \\ 1 & 2 \end{bmatrix}\begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 5 & 14 & 15 \\ 1 & 4 & 3 \end{bmatrix}.$$

It is to be noted that there is no definition of a matrix product ab when the number of columns of a is not the same as the number of rows of b. Thus the matrix product

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{bmatrix}$$

does not exist. In this case the matrix product

$$\begin{bmatrix} a_{11} & a_{12} & 0 \\ a_{21} & a_{22} & 0 \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{bmatrix}$$

does exist, as does the matrix product ba.

It follows that ba is not necessarily the same as ab for one of these products can exist without the other. Even though both products exist, and if a and b are square matrices with n rows and columns, they are not necessarily the same since, according to definition,

$$ba = \sum_{k=1}^{n} b_{ik}a_{kj}, \tag{3}$$

and the resulting matrix is not in general ab.

Thus multiplication is not commutative, and it is necessary to describe any product more precisely than in arithmetic and algebra. The product ab may be considered as b multiplied by a on the left (b premultiplied by a), or it can be considered as a multiplied by b on the right (a postmultiplied by b).

Premultiplication by a diagonal matrix gives a matrix in which each element of the original matrix is multiplied by the element in the same row of the diagonal matrix. Postmultiplication by a diagonal matrix produces a matrix in which each element of the original matrix is multiplied by the element in the same column of the original matrix. The premultiplication or postmultiplication of any matrix by a diagonal matrix with each of its (diagonal) elements identical results in the multiplication of every element in the original matrix by that diagonal element. In particular, premultiplication or postmultiplication by the diagonal matrix I does not change the value of the matrix. Thus

$$Ia = aI = a. \tag{4}$$

12.4 Basic multiplication laws. The fundamental associative and distributive laws of matrix algebra may be established.

$$
\begin{aligned}
a + b &= b + a \\
a(b + c) &= ab + ac \\
a + (b + c) &= (a + b) + c \\
(b + c)a &= ba + ca \\
(ab)c &= a(bc) \\
ab &\neq ba \text{ (in general).}
\end{aligned}
\tag{1}
$$

In addition to the products ab and ba discussed in the preceding section, there are other matrix products, where the definition is applicable, whose elements are composed of the sums of products of elements of a and b. Such products are $a'b$, ab', $b'a$, ba', $a'b'$, and $b'a'$. It is possible to write the elements of the ith row and the jth column of each of these

TABLE 12.4a

REPRESENTATIVE ELEMENTS IN DIFFERENT MATRIX PRODUCTS

Product	Element in the ith Row and the jth Column of the Product	Element in the jth Row and the ith Column in the Transpose of the Product
ab	$\sum_k a_{ik}b_{kj}$	$\sum_k a_{jk}b_{ki}$
ba	$\sum_k b_{ik}a_{kj}$	$\sum_k b_{jk}a_{ki}$
$a'b$	$\sum_k a_{ki}b_{kj}$	$\sum_k a_{kj}b_{ki}$
ab'	$\sum_k a_{ik}b_{jk}$	$\sum_k a_{jk}b_{ik}$
ba'	$\sum_k b_{ik}a_{jk}$	$\sum_k b_{jk}a_{ik}$
$b'a$	$\sum_k b_{ki}a_{kj}$	$\sum_k b_{kj}a_{ki}$
$a'b'$	$\sum_k a_{ki}b_{jk}$	$\sum_k a_{kj}b_{ik}$
$b'a'$	$\sum_k b_{ki}a_{jk}$	$\sum_k b_{kj}a_{ik}$

products since by definition the element in the jth row and the ith column of a' is identical to that of the jth row and the ith column of a. This is done in the second column of Table 12.4a.

A proper interpretation of Table 12.4a gives considerable information leading to the efficient computation of the different products that can be formed from a and b. Thus the elements of $a'b$ are formed by multiplying the elements a_{ki} of the ith column of a and the elements b_{kj} of the jth column of b. This means that $a'b$ can be computed with a column-by-column multiplication of a and b, with the column of a indicating the row and the column of b indicating the column of the product. If a and b (rather than a' and b') are available from previous computation, as they frequently are in least squares and correlation problems, the product $a'b$ is properly accomplished by a column-by-column multiplication.

Similarly the matrix ab' is produced by a row-by-row multiplication of a and b. In this case the ith row of a and the jth row of b are multiplied to get the element in the ith row and the jth column of the matrix ab'.

In a similar manner the matrix $a'b'$ can be obtained by multiplying columns of a by rows of b.

A computing form that aids in the calculation of a matrix product with the conventional method is obtained from placing the matrix a at the left of the product matrix, the matrix b above the product matrix such as

	b
a	ab

TABLE 12.4b

CALCULATION OF MATRIX PRODUCTS—ROW-BY-COLUMN MULTIPLICATION

				0.03846	0	0	0
				−0.01397	0.01912	0	0
		a	$b\rightarrow$	0.01577	−0.00553	0.02541	0
		$\downarrow$		−0.00282	−0.02268	0.01991	0.02322
1.00000	0.38462	−0.39338	−0.65416	0.02873	0.02437	−0.02302	−0.01519
0	1.00000	0.47720	0.18046	−0.00695	0.01239	0.01572	0.00419
0	0	1.00000	−0.87916	0.01825	0.01441	0.00791	−0.02041
0	0	0	1.00000	−0.00282	−0.02268	0.01991	0.02322
							$ab\uparrow$

In this case the place of each element in the product is easily located. An example is given in Table 12.4*b* in which an illustration from Table 13.7*b* is used.

The column-by-column and row-by-row multiplication are, in general, more easily accomplished than the conventional column-by-row type, since it is easier to associate the elements that are to be identified. As an illustration consider

$$a = \begin{vmatrix} 3 & 1 & 4 \\ 2 & 1 & 6 \end{vmatrix} \qquad b = \begin{vmatrix} 1 & 0 & 1 & 0 \\ 2 & 1 & 1 & -2 \end{vmatrix}.$$

To get $a'b$ we simply multiply the elements of columns and get

$$a'b = \begin{bmatrix} 7 & 2 & 5 & -4 \\ 3 & 1 & 2 & -2 \\ 16 & 6 & 10 & -12 \end{bmatrix}.$$

Similarly, if

$$a = \begin{bmatrix} 3 & 1 & 4 \\ 2 & 1 & 6 \end{bmatrix}$$

$$b = \begin{bmatrix} 1 & 0 & 1 \\ 2 & 1 & 1 \end{bmatrix}$$

then $a'b$ is obtained by multiplying elements of rows, and we have

$$ab' = \begin{bmatrix} 7 & 11 \\ 8 & 11 \end{bmatrix}.$$

A column-by-column multiplication could be used if the matrices were in mesh by rows. An illustration of this, taken from Table 13.7*a*, is presented in Table 12.4*c*. The odd-numbered rows compose the transpose of the matrix a, and the even-numbered rows those of matrix b.

An interchange of i and j in the elements of the second column of Table 12.4*a* gives the typical element in the transpose of the product. The identification of the term in the first row and the third column with the term of the last row and the second column of the table establishes the fact that

(2) $$(ab)' = b'a'.$$

Similar results hold for the pairs ba and $a'b'$, $a'b$ and $b'a$, ab' and $a'b$. In fact these results are special cases of the general statement (2). It

is easy to extend (2) since

$$(3) \qquad (abc)' = [a(bc)]' = (bc)'a' = c'b'a', \text{ etc.}$$

The fact that $a'b$ is a column-by-column multiplication and that ab'

TABLE 12.4c

CALCULATION OF MATRIX PRODUCT—COLUMN-BY-COLUMN MULTIPLICATION

1.00000	0	0	0
0.03846	0	0	0
0.38462	1.00000	0	0
−0.01397	0.01912	0	0
−0.39338	0.47720	1.00000	0
0.01577	−0.00553	0.02541	0
−0.65416	0.18046	−0.87916	1.00000
−0.00282	−0.02268	0.01991	0.02322
0.02873	0.02437	−0.02302	−0.01519
−0.00695	0.01239	0.01572	0.00419
0.01825	0.01441	0.00791	−0.02041
−0.00282	−0.02268	0.01991	0.02322

is a row-by-row multiplication is useful in extending the techniques of actual matrix multiplication. Thus

$$(4) \qquad ab = (a')'b$$

so that ab is the column-by-column multiplication of a' and b. The product ab can be computed by taking the transpose of a and using a column-by-column multiplication. In a similar fashion

$$(5) \qquad ab = a(b')',$$

and the product can be computed by a row-by-row multiplication of the elements of a and the elements of b'.

The concept of column-by-column multiplication of $a'b$ and of row-by-row multiplication of ab' seems to the writer to be just as important, from a manipulative standpoint, as the conventional row-by-column multiplication. At least we should recognize these operations as specific matrix operations.

In some situations the matrices a' and b are in mesh, that is, the first row of a is followed by the first row of b, then by the second row of a,

the second row of b, etc. Thus we may have the computational form

a_{11}	a_{12}	a_{13}
b_{11}	b_{12}	b_{13}
a_{21}	a_{22}	a_{23}
b_{21}	b_{22}	b_{23}
a_{31}	a_{32}	a_{33}
b_{31}	b_{32}	b_{33}

It is easy to form the column-by-column multiplication that is the value of $a'b$, since all elements to be multiplied are in adjacent rows.

As an illustration of this we note that (11.4.2) can be interpreted as a matrix multiplication. If a is the column vector composed of g_{i1}, $g_{i2\cdot 1}$, $\cdots$, etc., and b is the column vector composed of $b_{1,\,p+1}$, $b_{2,\,p+1\cdot 1}$, etc., then (11.4.2) may be written

$$a_{i,\,p+1} = a'b \tag{6}$$

and the computational technique used in getting results from the entries of Table 11.4a, Table 11.5a, and Table 11.5b is precisely and simply described.

12.5 The rank of a matrix. A matrix is a rectangular array of elements and a determinant is a specified function of the elements of a square matrix. In general, there is a determinant for each square matrix, and there is one for each square array of elements that can be obtained from any matrix by deletion of one or more rows and columns. These determinants are called determinants of the matrix.

For the square matrix there is a determinant whose order is equal to that of the matrix. If this determinant is zero, the matrix is said to be singular; otherwise it is non-singular. It is possible to calculate the determinants of the matrix of all orders and to determine the value of each. The *rank* of the matrix is defined to be the order of the determinant (or series of determinants) of highest order that is not zero. Thus to state that a 20 by 15 matrix is of rank 5 means that every determinant of order 6 or more is zero.

12.6 Summary. The basic manipulative rules of introductory matrix algebra are presented in this chapter with the appropriate definitions. Similar material may be found in [A]. The concept of the inverse matrix, its relation to the problem of solving simultaneous linear

equations, and a variety of direct methods of calculation are presented in Chapters 13 and 14.

REFERENCES

A. Other presentations of the elements of matrix theory needed for linear problems may be found in

1. D. B. Duncan and J. F. Kenney, *On the Solution of Normal Equations and Related Topics*, Edwards Brothers, Ann Arbor, 1946. See pp. 1–14.
2. L. L. Thurstone, *Multiple Factor Analysis*, University of Chicago Press, Chicago, 1947. See pp. 1–50.
3. The reader is also referred to the work of Banachiewicz, who has used the concept of the cracovian in handling linear problems. A cracovian is similar to a matrix, but the multiplications are column by column. The reader is referred to the references given in Chapter 6 to the work of Banachiewicz.

EXERCISES

1. $$\begin{bmatrix} 2 & 1 \\ 3 & 6 \end{bmatrix} + \begin{bmatrix} 4 & 2 \\ 9 & 7 \end{bmatrix}$$

2. $$\begin{bmatrix} 2 & 4 & 6 \\ 3 & 1 & 2 \end{bmatrix} + \begin{bmatrix} 9 & 3 & 2 \\ 1 & 2 & 7 \end{bmatrix}$$

3. $$\begin{bmatrix} 2 & 1 \\ 3 & 6 \end{bmatrix} - \begin{bmatrix} 4 & 2 \\ 9 & 7 \end{bmatrix}$$

4. $$\begin{bmatrix} 2 & 4 & 6 \\ 3 & 1 & 2 \end{bmatrix} - \begin{bmatrix} 9 & 3 & 2 \\ 1 & 2 & 7 \end{bmatrix}$$

5. $$\begin{bmatrix} 3 & 1 \\ 3 & 6 \end{bmatrix} \begin{bmatrix} 4 & 2 \\ 9 & 7 \end{bmatrix}$$

6. $$\begin{bmatrix} 3 & 1 \\ 3 & 6 \end{bmatrix} \begin{bmatrix} 4 & 9 \\ 2 & 7 \end{bmatrix}$$

7. $$\begin{bmatrix} 3 & 3 \\ 1 & 6 \end{bmatrix} \begin{bmatrix} 4 & 2 \\ 9 & 7 \end{bmatrix}$$

8. $$\begin{bmatrix} 3 & 3 \\ 1 & 6 \end{bmatrix} \begin{bmatrix} 4 & 9 \\ 2 & 7 \end{bmatrix}$$

9. $$\begin{bmatrix} 2 & 4 & 6 \\ 3 & 1 & 2 \end{bmatrix} \begin{bmatrix} 9 & 1 \\ 3 & 2 \\ 2 & 7 \end{bmatrix}$$

10. $$a = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & 4 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad b = \begin{bmatrix} 6 & 0 & 0 \\ 7 & 8 & 0 \\ 7 & 4 & 2 \end{bmatrix},$$

calculate ab, ba, $a'b$, ba', $b'a$, ab', $a'b'$, and $b'a'$.

11. Calculate a^2, if

$$a = \begin{bmatrix} 1.0 & 0.4 & 0.5 & 0.6 \\ 0.4 & 1.0 & 0.3 & 0.4 \\ 0.5 & 0.3 & 1.0 & 0.2 \\ 0.6 & 0.4 & 0.2 & 1.0 \end{bmatrix}.$$

12. Calculate ab' and $a'b$, if

$$a = \begin{bmatrix} -1.217 & 0.312 & -0.413 \\ 0.412 & 1.477 & 0.232 \\ -0.913 & -0.822 & 1.111 \end{bmatrix}$$

$$b = \begin{bmatrix} 862 & 823 & 332 \\ 114 & 923 & -611 \\ 444 & -919 & 1027 \end{bmatrix}.$$

CHAPTER 13

The Inverse Matrix and Its Calculation with Approximate Methods

13.1 Introduction. Generally speaking, one matrix cannot be divided by another. However, there are matrix operations that play roles corresponding to division in algebra and arithmetic if the "denominator" is a square non-singular matrix. The result of "dividing" the identity matrix by the square non-singular matrix is the inverse of a or the reciprocal of a. It is indicated by a^{-1} and is any matrix that satisfies the matrix equation

$$a^{-1}a = aa^{-1} = I. \tag{1}$$

We may proceed with a generalized "division" by premultiplying or postmultiplying by a^{-1} if the number of rows and columns of matrices permit multiplication. Thus $a^{-1}b$ and ba^{-1} are matrix generalizations of the quotient b/a.

13.2 The adjugate or adjoint. The inverse matrix can be approached by various methods, but it is perhaps best in this book to relate it to the solution of the simultaneous equations (4.1.1). These equations can be expressed in the matrix form

$$ax = f, \tag{1}$$

where a is the p by p matrix $[a_{ij}]$, x is the p by 1 matrix $[x_i]$, and f is the p by 1 matrix $[a_{i,\,p+1}]$.

For simplicity of presentation we limit the discussion to (4.1.2), a special case of (4.1.1), with $p = 4$. Results are applicable to the more general case.

We first define a matrix, called the *adjugate* or *adjoint* of a, in which each element is the cofactor of an element of a. The cofactor of a_{ij} is conventionally placed in the ij position in the new matrix. Thus the matrix

$$a = \begin{bmatrix} 2 & 4 & 6 \\ 1 & 3 & 5 \\ 2 & 1 & 4 \end{bmatrix}$$

of Table 10.12*c* would yield the matrix

$$\text{adj } a = \begin{bmatrix} 7 & 6 & -5 \\ -10 & -4 & 6 \\ 2 & -4 & 2 \end{bmatrix}$$

when the elements of a are replaced by their cofactors. It can be shown by the principles of determinants that row-by-row multiplication of a and adj a gives the value of the determinant when the rows correspond, and zero otherwise. A similar statement holds for column-by-column multiplication. The reader unfamiliar with this fact should verify this statement as applied to the numerical illustration just given where the value of the determinant is 8. For purposes of linear computation it seems preferable to follow Aitken [A] and use a matrix that is the transpose of the matrix just written. If we let A_{ij} be the cofactor of a_{ji} in a, we can define

$$A = [A_{ij}]. \tag{2}$$

This A is the transpose of adj a defined above. In the foregoing numerical illustration the value of A is

$$A = \begin{bmatrix} 7 & -10 & 2 \\ 6 & -4 & -4 \\ -5 & 6 & 2 \end{bmatrix}.$$

When the matrix of cofactors appears in this form we have the basic statement

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} & A_{13} & A_{14} \\ A_{21} & A_{22} & A_{23} & A_{24} \\ A_{31} & A_{32} & A_{33} & A_{34} \\ A_{41} & A_{42} & A_{43} & A_{44} \end{bmatrix} = \begin{bmatrix} \Delta & 0 & 0 & 0 \\ 0 & \Delta & 0 & 0 \\ 0 & 0 & \Delta & 0 \\ 0 & 0 & 0 & \Delta \end{bmatrix} \tag{3}$$

which appears in the general form

$$aA = \Delta I, \tag{4}$$

where ΔI is a diagonal matrix. The terms "adjugate" and "adjoint" have been used to denote each of the matrices A and adj a. In the remaining pages of this book we call A the adjoint matrix and $A' = \text{adj } a$, the adjugate matrix. The A matrix, rather than the A' matrix, is utilized in the theorems that follow. It should be noted that A' is symmetric if a is symmetric. In this case, $A = A'$.

We can show in a similar fashion that

$$\begin{bmatrix} A_{11} & A_{12} & A_{13} & A_{14} \\ A_{21} & A_{22} & A_{23} & A_{24} \\ A_{31} & A_{32} & A_{33} & A_{34} \\ A_{41} & A_{42} & A_{43} & A_{44} \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} = \begin{bmatrix} \Delta & 0 & 0 & 0 \\ 0 & \Delta & 0 & 0 \\ 0 & 0 & \Delta & 0 \\ 0 & 0 & 0 & \Delta \end{bmatrix}$$

and in general that

$$Aa = \Delta I = aA. \tag{5}$$

The adjoint (adjugate) exists for every square matrix with finite elements no matter whether the matrix is singular or not. If the matrix is singular, (5) becomes

$$Aa = aA = 0, \tag{6}$$

and no inverse exists.

13.3 The inverse matrix and the solution of simultaneous equations. If a is non-singular, $\Delta \neq 0$, and we may divide every element A_{ij} and every element in ΔI in (13.2.5) by Δ to get

$$ca = I = ac \quad \text{where } c_{ij} = \frac{A_{ij}}{\Delta}. \tag{1}$$

Now (1) is of the same form as (13.1.1), with $c = a^{-1}$. It follows that an inverse matrix is an adjoint matrix, with each element divided by the determinant of the matrix. This fact can be used in the direct calculation of the inverse. Thus the inverse of

$$a = \begin{bmatrix} 2 & 4 & 6 \\ 1 & 3 & 5 \\ 2 & 1 & 4 \end{bmatrix}$$

is

$$a^{-1} = \left[\frac{A}{8}\right] = \begin{bmatrix} \frac{7}{8} & -\frac{10}{8} & \frac{2}{8} \\ \frac{6}{8} & -\frac{4}{8} & -\frac{4}{8} \\ -\frac{5}{8} & \frac{6}{8} & \frac{2}{8} \end{bmatrix}.$$

The transpose of the inverse, $(a^{-1})' = c'$, is also useful in computation. It can be demonstrated in general that

$$(a^{-1})' = (a')^{-1}. \tag{2}$$

If a is symmetric, A is symmetric with $c' = c$ [A]. The value of a^{-1} in the matrix equations (13.1.1) can be proved to be unique [A]. The inverse of a diagonal non-singular matrix is also a diagonal non-singular

matrix, with elements that are the reciprocals of the elements of the diagonal matrix.

We are now in a position to use the inverse matrix in solving simultaneous equations. If we premultiply (13.2.1) by $c = a^{-1}$, we get

$$a^{-1}ax = a^{-1}f \quad \text{or} \quad x = a^{-1}f = cf. \tag{3}$$

Thus premultiplication of both sides by a^{-1} yields the solution that, no matter how obtained, is the equivalent of a premultiplication by a^{-1}. As a simple illustration we take the equations

$$\begin{aligned} x_1 + x_2 + x_3 &= 0 \\ x_1 - x_2 + x_3 &= -4 \\ x_1 + x_2 - x_3 &= 6, \end{aligned}$$

where

$$a = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \end{bmatrix}, \quad f = \begin{bmatrix} 0 \\ -4 \\ 6 \end{bmatrix}, \quad A = \begin{bmatrix} 0 & 2 & 2 \\ 2 & -2 & 0 \\ 2 & 0 & -2 \end{bmatrix}, \quad \Delta = 4$$

$$c = a^{-1} = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & 0 \\ \frac{1}{2} & 0 & -\frac{1}{2} \end{bmatrix}$$

so that

$$x = cf = [1, 2, -3]. \tag{4}$$

In general the solution (3) may be expanded in the form

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & c_{13} & c_{14} \\ c_{21} & c_{22} & c_{23} & c_{24} \\ c_{31} & c_{32} & c_{33} & c_{34} \\ c_{41} & c_{42} & c_{43} & c_{44} \end{bmatrix} \begin{bmatrix} a_{15} \\ a_{25} \\ a_{35} \\ a_{45} \end{bmatrix} \tag{5}$$

$$= \begin{bmatrix} \frac{A_{11}}{\Delta} & \frac{A_{12}}{\Delta} & \frac{A_{13}}{\Delta} & \frac{A_{14}}{\Delta} \\ \frac{A_{21}}{\Delta} & \frac{A_{22}}{\Delta} & \frac{A_{23}}{\Delta} & \frac{A_{24}}{\Delta} \\ \frac{A_{31}}{\Delta} & \frac{A_{32}}{\Delta} & \frac{A_{33}}{\Delta} & \frac{A_{34}}{\Delta} \\ \frac{A_{41}}{\Delta} & \frac{A_{42}}{\Delta} & \frac{A_{43}}{\Delta} & \frac{A_{44}}{\Delta} \end{bmatrix} \begin{bmatrix} a_{15} \\ a_{25} \\ a_{35} \\ a_{45} \end{bmatrix}.$$

Thus

$$x_1 = \frac{A_{11}a_{15} + A_{12}a_{25} + A_{13}a_{35} + A_{14}a_{45}}{\Delta}$$

$$(6) \qquad = \begin{bmatrix} a_{15} & a_{12} & a_{13} & a_{14} \\ a_{25} & a_{22} & a_{23} & a_{24} \\ a_{35} & a_{32} & a_{33} & a_{34} \\ a_{45} & a_{42} & a_{43} & a_{44} \end{bmatrix} \div \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}.$$

Similar statements can be made for the other values of x_i and for the more general case of p variables. Thus Cramer's rule (10.9.1) can be derived from (3).

It is not wise to evaluate, as is conventionally advised, all the different determinantal ratios that are the values of x_i by using alternative forms of (6). It is better to compute one of them by condensation processes and then to use the elimination equations as the basis of the back solution (as is demonstrated in Chapters 4 to 6) or to use the substitute for the back solution described in Chapter 11.

13.4 Additional definitions and theorems. The inverse of a product is needed in many applications. This can be obtained if the inverse of each factor exists and is known. Thus

$$(b^{-1}a^{-1})(ab) = b^{-1}(a^{-1}a)b = b^{-1}Ib = b^{-1}b = I$$

so that

$$(1) \qquad (b^{-1}a^{-1})^{-1} = ab \qquad \text{and} \qquad (ab)^{-1} = b^{-1}a^{-1}.$$

The same type of reasoning can be extended to show that the inverse of a product is the product of the inverses in reverse order.

A square matrix of special interest is the so-called *orthogonal* matrix. An orthogonal matrix is the one in which $a'a = aa' = I$ so that $a' = a^{-1}$. Application of section 12.4 shows that the row-by-row multiplication of a by a, or the column-by-column multiplication of a by a, is zero when the two rows (or columns) are different but that unity results when the two rows (or columns) are the same.

Of special interest in elimination methods is the *triangular matrix* in which all elements above (or below) the main diagonal are zero. A triangular matrix is a square matrix that obtains its name from the fact that all the elements that need to be recorded (the others are zero) appear in triangular form. For example, the matrix of the coefficients of the elimination equations of the square root method, section 6.5, ap-

pears as

$$s = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ 0 & s_{22\cdot 1} & \cdots & s_{2p\cdot 1} \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & & s_{pp\cdot(p-1)} \end{bmatrix}. \tag{2}$$

The matrix of the coefficients of the first doublet rows of the elimina equations of a Gauss-Doolittle solution is

$$g' = \begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1p} \\ 0 & g_{22\cdot 1} & \cdots & g_{2p\cdot 1} \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & & g_{pp\cdot(p-1)} \end{bmatrix} \tag{3}$$

and the matrix of the second of the doublet rows is

$$b = \begin{bmatrix} 1 & b_{12} & \cdots & b_{1p} \\ 0 & 1 & \cdots & b_{2p\cdot 1} \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & 1 \end{bmatrix}. \tag{4}$$

Similarly the elimination form of the compact method of single sion contains the two triangular matrices

$$g = \begin{bmatrix} g_{11} & 0 & \cdots & 0 \\ g_{21} & g_{22\cdot 1} & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ g_{p1} & g_{p2\cdot 1} & \cdots & g_{pp\cdot(p-1)} \end{bmatrix} \tag{5}$$

and the b of (4), which are in mesh along the main diagonal.

The basic idea of the forward solution of the elimination proced using division methods may now be stated. It is desired that the efficients of the elimination equations have the form of one or m triangular matrices such as g and b, where g, b, and a are related $gb = a$. This is proved by reducing an additional row and colum the original matrix to zero at the elimination of each variable until rows and columns of the original matrix have become zero [B]. compact method of single division shows in conventional matrix (r by-column) form that $a - gb = 0$, whereas the Gauss-Doolittle met establishes the corresponding statement (for symmetric matrices) using column-by-column multiplication. The square root method u

column-by-column multiplication of the triangular matrix s to show that $a - s's = 0$. A more formal proof of these relations is presented elsewhere. In any case we have the relations

$$a = gb \quad \text{and} \quad a = s's. \tag{6}$$

If a is not symmetric, g refers to the matrix of (5) or to the transpose of the top rows of the first four columns of the elimination equations of Table 6.4g. If a is symmetric, g refers to the transpose of g' in (3). Here multiplication by the diagonal matrix

$$G = \begin{bmatrix} g_{11} & 0 & \cdots & 0 \\ 0 & g_{22\cdot 1} & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & g_{pp\cdot(p-1)} \end{bmatrix} = G' \tag{7}$$

gives

$$g' = Gg, \qquad g = b'G, \qquad \text{and} \qquad b' = gG^{-1}. \tag{8}$$

It follows from (6), if a is symmetric, that

$$\begin{aligned} a &= (b'G)b = b'Gb = (b'G^{\frac{1}{2}})(G^{\frac{1}{2}}b) = g'G^{-1}g \\ &= (g'G^{-\frac{1}{2}})'(G^{-\frac{1}{2}}g) = (G^{-\frac{1}{2}}g)'(G^{-\frac{1}{2}}g) = s's. \end{aligned} \tag{9}$$

The formula shows why the square root method features division by the square root of diagonal terms.

A special technique is useful in the determination of the inverse of a triangular matrix. Suppose a is a triangular matrix. Then its inverse is the solution of $ax = I$. The synthetic solution takes the form, when $p = 4$,

$$\begin{array}{cccccccc} a_{11} & a_{12} & a_{13} & a_{14} & 1 & 0 & 0 & 0 \\ 0 & a_{22} & a_{23} & a_{24} & 0 & 1 & 0 & 0 \\ 0 & 0 & a_{33} & a_{34} & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & a_{44} & 0 & 0 & 0 & 1 \end{array} \tag{10}$$

and the inverse can be computed by a conventional back solution. The identity matrix need not be recorded as it has unity for each diagonal term and zero elsewhere. The computation of the inverse of the triangular matrix of Table 13.8a is given in Table 13.4a.

pears as

$$s = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ 0 & s_{22\cdot 1} & \cdots & s_{2p\cdot 1} \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & & s_{pp\cdot(p-1)} \end{bmatrix}. \tag{2}$$

The matrix of the coefficients of the first doublet rows of the elimination equations of a Gauss-Doolittle solution is

$$g' = \begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1p} \\ 0 & g_{22\cdot 1} & \cdots & g_{2p\cdot 1} \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & & g_{pp\cdot(p-1)} \end{bmatrix} \tag{3}$$

and the matrix of the second of the doublet rows is

$$b = \begin{bmatrix} 1 & b_{12} & \cdots & b_{1p} \\ 0 & 1 & \cdots & b_{2p\cdot 1} \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & 1 \end{bmatrix}. \tag{4}$$

Similarly the elimination form of the compact method of single division contains the two triangular matrices

$$g = \begin{bmatrix} g_{11} & 0 & \cdots & 0 \\ g_{21} & g_{22\cdot 1} & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ g_{p1} & g_{p2\cdot 1} & \cdots & g_{pp\cdot(p-1)} \end{bmatrix} \tag{5}$$

and the b of (4), which are in mesh along the main diagonal.

The basic idea of the forward solution of the elimination procedures using division methods may now be stated. It is desired that the coefficients of the elimination equations have the form of one or more triangular matrices such as g and b, where g, b, and a are related by $gb = a$. This is proved by reducing an additional row and column of the original matrix to zero at the elimination of each variable until all rows and columns of the original matrix have become zero [B]. The compact method of single division shows in conventional matrix (row-by-column) form that $a - gb = 0$, whereas the Gauss-Doolittle method establishes the corresponding statement (for symmetric matrices) by using column-by-column multiplication. The square root method uses

column-by-column multiplication of the triangular matrix s to show that $a - s's = 0$. A more formal proof of these relations is presented elsewhere. In any case we have the relations

$$a = gb \quad \text{and} \quad a = s's. \tag{6}$$

If a is not symmetric, g refers to the matrix of (5) or to the transpose of the top rows of the first four columns of the elimination equations of Table 6.4g. If a is symmetric, g refers to the transpose of g' in (3). Here multiplication by the diagonal matrix

$$G = \begin{bmatrix} g_{11} & 0 & \cdots & 0 \\ 0 & g_{22\cdot 1} & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & g_{pp\cdot(p-1)} \end{bmatrix} = G' \tag{7}$$

gives

$$g' = Gg, \quad g = b'G, \quad \text{and} \quad b' = gG^{-1}. \tag{8}$$

It follows from (6), if a is symmetric, that

$$\begin{aligned} a &= (b'G)b = b'Gb = (b'G^{1/2})(G^{1/2}b) = g'G^{-1}g \\ &= (g'G^{-1/2})'(G^{-1/2}g) = (G^{-1/2}g)'(G^{-1/2}g) = s's. \end{aligned} \tag{9}$$

The formula shows why the square root method features division by the square root of diagonal terms.

A special technique is useful in the determination of the inverse of a triangular matrix. Suppose a is a triangular matrix. Then its inverse is the solution of $ax = I$. The synthetic solution takes the form, when $p = 4$,

$$\begin{array}{cccccccc} a_{11} & a_{12} & a_{13} & a_{14} & 1 & 0 & 0 & 0 \\ 0 & a_{22} & a_{23} & a_{24} & 0 & 1 & 0 & 0 \\ 0 & 0 & a_{33} & a_{34} & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & a_{44} & 0 & 0 & 0 & 1 \end{array} \tag{10}$$

and the inverse can be computed by a conventional back solution. The identity matrix need not be recorded as it has unity for each diagonal term and zero elsewhere. The computation of the inverse of the triangular matrix of Table 13.8a is given in Table 13.4a.

TABLE 13.4*a*

INVERSE OF TRIANGULAR MATRIX

1.000	0.313	0.280	0.495	
0	0.950	0.594	0.521	
0	0	0.754	0.471	a
0	0	0	0.512	
1.000	0	0	0	
−0.330	1.053	0	0	a^{-1}
−0.112	−0.829	1.326	0	
−0.529	−0.308	−1.220	1.953	

13.5 The calculation of the inverse matrix with pivotal methods. The inverse matrix can be obtained by solving the matrix equation

$$ax = I \tag{1}$$

by the various pivotal methods. There are a number of variations of these which feature (*a*) the formal reduction of the identity matrix, (*b*) the formal reduction of the identity matrix without a back solution, and (*c*) the inverse matrix without the reduction of the identity matrix. In each case an easier technique is available if a, and hence a^{-1}, is symmetric. In addition there is the possibility of using each of the methods outlined in Chapters 4 to 6.

The following sections of this chapter deal with approximate methods only. Exact methods are presented in the following chapter. In the interest of brevity, only the Gauss-Doolittle and square root methods are used for the symmetric case and the generalized Doolittle and compact method of single division for the general case in which symmetry is not assumed.

13.6 The inverse matrix with solution of I by approximate methods. The Gauss-Doolittle presentation is given in Table 13.6*a*. The forward solution is as usual. The back solution starts with the values of x_4, which are placed in the last row. The values of the last column are then written from symmetry. The entries of the third row are then computed by the back solution and are repeated in the third column, etc. This technique gives the value of $(a^{-1})'$, which can be used to accomplish a row-by-row multiplication by a, since $(a^{-1})' = a^{-1}$ when a is symmetric.

The corresponding square root solution is given in Table 13.6*b*. The illustration is one previously used [C].

TABLE 13.6a

INVERSE MATRIX REDUCTION OF I—GAUSS-DOOLITTLE METHOD

1.0000	0.4000	0.5000	0.6000	1	0	0	0
*	1.0000	0.3000	0.4000	0	1	0	0
*	*	1.0000	0.2000	0	0	1	0
*	*	*	1.0000	0	0	0	1
1.0000	0.4000	0.5000	0.6000	1.0000	0	0	0
1.0000	0.4000	0.5000	0.6000	1.0000	0	0	0
	0.8400	0.1000	0.1600	−0.4000	1.0000	0	0
	1.0000	0.1190	0.1905	−0.4762	1.1905	0	0
		0.7381	−0.1190	−0.4524	−0.1190	1.0000	0
		1.0000	−0.1612	−0.6129	−0.1612	1.3548	0
			0.5903	−0.5966	−0.2097	0.1612	1.0000
			1.0000	−1.0107	−0.3552	0.2731	1.6941
2.0708	−0.1913	−0.7759	−1.0107				
−0.1913	1.2842	−0.2185	−0.3552				
−0.7759	−0.2185	1.3988	0.2731				
−1.0107	−0.3552	0.2731	1.6941				

TABLE 13.6b

INVERSE REDUCTION OF I—SQUARE ROOT METHOD

1.0000	0.313	0.280	0.495	1	0	0	0
*	1.000	0.652	0.650	0	1	0	0
*	*	1.000	0.803	0	0	1	0
*	*	*	1.000	0	0	0	1
1.000	0.313	0.280	0.495	1.000	0	0	0
	0.950	0.594	0.521	−0.329	1.053	0	0
		0.754	0.471	−0.112	−0.830	1.326	0
			0.512	−0.529	−0.308	−1.220	1.953
1.401	−0.091	0.497	−1.033				
−0.091	1.891	−0.723	−0.602				
0.497	−0.723	3.247	−2.383				
−1.033	−0.602	−2.383	3.814				

The method may be applied similarly to non-symmetric matrices with the use of a generalized Doolittle or compact method of single division, although the back solution is somewhat more complicated than for the symmetric case. The solution using the compact method of single division is indicated in Table 13.6*c*, where the back solution,

TABLE 13.6*c*

INVERSE MATRIX REDUCTION OF I—COMPACT METHOD OF SINGLE DIVISION

26	−10	15	32	1	0	0	0
19	45	−14	−8	0	1	0	0
−12	16	27	13	0	0	1	0
32	29	−35	28	0	0	0	1
26.00000	−0.38462	0.57692	1.23077	0.03846	0	0	0
19.00000	52.30778	−0.47720	−0.60000	−0.01397	0.01912	0	0
−12.00000	11.38456	39.35575	0.87916	0.01577	−0.00553	0.02541	0
32.00000	41.30784	−33.74934	43.07113	−0.00282	−0.02268	0.01191	0.02322
0.02873	−0.00695	0.01825	−0.00282				
0.02487	0.01239	0.01441	−0.02268				
−0.02302	0.01572	0.00791	0.01991				
−0.01519	0.00419	−0.02041	0.02322				

obtained by treating each right-hand column separately, is the transpose of the inverse. This is the form useful in checking because multiplication by rows of c' and a is possible.

13.7 The inverse matrix without a back solution with approximate methods. The forward part of this solution is identical with that of the last section, but the inverse is calculated from the columns of the elimination equations that are headed by the identity matrix. (A slightly different technique is necessary for the compact method of single division.) The values of the inverse, or of its transpose, are obtained by multiplying, by columns, the elements in the last rows of the right half of the table. This is illustrated in Table 13.6*b*, where

$$1.401 = (1.000)^2 + (-0.329)^2 + (-0.112)^2 + (-0.529)^2$$

$$-0.091 = (1.000)(0) + (-0.329)(1.053) + (-0.112)(-0.830) + (-0.529)(-0.308),$$

etc.

It is similarly illustrated in Table 13.6*a*, where the column-by-column multiplication utilizes entries from each of the couplet rows. Thus

$$2.0707 = (1.0000)(1.0000) + (-0.4000)(-0.4762) + (-0.4524)(-0.6129) + (-0.5966)(-1.0107)$$

and

$$-0.1914 = (1.0000)(0) + (-0.4000)(1.1905) + (-0.4524)(-0.1612) + (-0.5966)(-0.3552).$$

$$-0.1913 = (0)(1.0000) + (1.0000)(-0.4762) + (-0.1190)(-0.6129) + (-0.2097)(-1.0107),$$

etc.

A similar technique holds for non-symmetric matrices. A generalized Doolittle presentation for a non-symmetric matrix is given in Table 13.7*a*; the corresponding compact method of single division is given in Table 13.7*b*.

TABLE 13.7*a*

INVERSE MATRIX WITHOUT BACK SOLUTION—GENERALIZED DOOLITTLE METHOD

26	−10	15	32	1	0	0	0
19	45	−14	−8	0	1	0	0
−12	16	27	13	0	0	1	0
32	29	−35	28	0	0	0	1
26.00000	19.00000	−12.00000	32.00000	1.00000	0	0	0
1.00000	−0.38462	0.57692	1.23077	0.03846	0	0	0
	52.30778	11.38456	41.30784	0.38462	1.00000	0	0
	1.00000	−0.47720	−0.60000	−0.01397	0.01912	0	0
		39.35575	−33.74934	−0.39338	0.47720	1.00000	0
		1.00000	0.87916	0.01577	−0.00553	0.02541	0
			43.07113	−0.65416	0.18046	−0.87916	1
			1.00000	−0.00282	−0.02268	0.01991	0.02322
0.02873	−0.00695	0.01825	−0.00282	0.02873	0.02437	−0.02302	−0.01519
0.02437	0.01239	0.01441	−0.02268	−0.00695	0.01239	0.01572	0.00419
−0.02302	0.01572	0.00791	0.01991	0.01825	0.01441	0.00791	−0.02041
−0.01519	0.00419	−0.02041	0.02322	−0.00282	−0.02268	0.01991	0.02322

The inverse of a is written on the right of the last four rows in Tables 13.7*a* and 13.7*b*. The transpose of the inverse is written on the left of the last four rows in Table 13.7*a*. The transpose of the inverse is useful in checking since the value of $a^{-1}a = ca$ can then be computed by a row-by-row multiplication of c' and a. It is the transpose of c, rather than c itself, that is useful in the solution of linear equations, and it may be feasible to record only c'. Thus the c matrix could be omitted in

TABLE 13.7*b*

INVERSE MATRIX WITHOUT BACK SOLUTION—COMPACT METHOD OF SINGLE DIVISION

26	−10	15	32	1	0	0	0
19	45	−14	−8	0	1	0	0
−12	16	27	13	0	0	1	0
32	29	−35	28	0	0	0	1
1	0	0	0				
0	1	0	0				
0	0	1	0				
0	0	0	1				
26.00000	−0.38462	0.57692	1.23077	0.03846	0	0	0
19.00000	52.30778	−0.47720	−0.60000	−0.01397	0.01912	0	0
−12.00000	11.38456	39.35575	0.87916	0.01577	−0.00553	0.02541	0
32.00000	41.30784	−33.74934	43.07113	−0.00282	−0.02268	0.01991	0.02322
1.00000	0.38462	−0.39338	−0.65416	0.02873	0.02437	−0.02302	−0.01519
0	1.00000	0.47720	0.18046	−0.00695	0.01239	0.01572	0.00419
0	0	1.00000	−0.87916	0.01825	0.01441	0.00791	−0.02041
0	0	0	1.00000	−0.00282	−0.02268	0.01991	0.02322

Table 13.7*a* and only the c' matrix recorded. Similarly the recording of the c matrix could be omitted from Table 13.7*b*, and the results could be given in the c' matrix, which could be inserted in the first four columns at the bottom.

The values of c_{ij} and c'_{ij} in Tables 13.7*a* and 13.7*b* are computed as follows:

$$\begin{aligned} c_{11} = c'_{11} &= (1.00000)(0.03846) + (0.38462)(-0.01397) \\ &\quad + (-0.39338)(0.01577) + (-0.65416)(-0.00282) \\ &= 0.02873 \end{aligned}$$

$$\begin{aligned} c_{12} = c'_{21} &= (1.00000)(0) + (0.38462)(0.01912) \\ &\quad + (-0.39338)(-0.00553) \\ &\quad + (-0.65416)(-0.02268) \\ &= 0.02437 \end{aligned}$$

$$\begin{aligned} c_{21} &= c'_{12}(0)(0.03846) + (1.00000)(-0.01397) + (0.47720)(0.01577) \\ &\quad + (0.18046)(-0.00282) \\ &= -0.00695 \end{aligned}$$

$$\begin{aligned} c_{22} &= c'_{22}(0)(0) + (1.00000)(0.01912) + (0.47720)(-0.00553) \\ &\quad + (0.18046)(-0.02268) \\ &= 0.01239. \end{aligned}$$

This operation in Table 13.7*a* may be described as a premultiplication of the elements of the matrix of second rows by the transpose of the matrix of first rows and, indeed, this is formally indicated in 13.7*b*, which includes the illustration of Table 12.4*b*.

A matrix proof is presented in justification of the techniques of this section. The basic matrix equation is

$$ax = I. \tag{1}$$

The matrix a is factored into gb by the generalized Doolittle method so that

$$a = gb \quad \text{and} \quad a' = b'g' \tag{2}$$

and

$$a^{-1} = b^{-1}g^{-1}. \tag{3}$$

The linear operations that transform a into b and a' into g' are also useful to transform the right-hand side. Thus the synthetic matrix form of the generalized Doolittle method is

(4)

a	I
g'	$(b^{-1})'$
b	g^{-1}
c'	c

where the matrices involving g' contain the top of the doublet rows of the elimination equations and the matrices involving b's contain the bottom rows. The terms on the right are determined by the equations

$$\begin{aligned} gbx &= I \quad \text{so that} \quad bx = (g)^{-1} \quad \text{and} \\ b'g'x &= I \quad \text{so that} \quad g'x = (b')^{-1}. \end{aligned} \tag{5}$$

The value of a^{-1} is accomplished by the column-by-column multiplication of $(b^{-1})'$ and $(g)^{-1}$.

A similar argument holds for the solution of Table 13.7*b*. The syn-

thetic matrix form is

(6)

a	I
I	
b / g	g^{-1}
b^{-1}	a^{-1}

so that $a^{-1} = b^{-1}(g)^{-1}$ is accomplished by a row-by-column multiplication.

The synthetic matrix forms for the symmetric methods are similar. The Gauss-Doolittle is a special case of the generalized Doolittle, with $a' = a$. The square root solution is described by the synthetic matrix form

(7)

a	I
s	sa^{-1}
c'	c

where $(sa^{-1})'(sa^{-1}) = a^{-1}s'sa^{-1} = a^{-1}$.

13.8 The inverse matrix without reduction of I with approximate methods. The formal reduction of the identity matrix can be eliminated. It is necessary only to get the values of the triangular matrices g and b and then to obtain $(g)^{-1}$ and $(b')^{-1}$. The value of a^{-1} is then determined by (13.7.3).

This method is especially applicable to the square root method since only one triangular matrix need be inverted. The synthetic matrix form of this solution appears as

(1)

a
s
s^{-1}
a^{-1}

where s^{-1} is the inverse of s and $a^{-1} = (s^{-1})(s^{-1})'$ is obtained by multiplying s^{-1} by columns. The method is applied to the illustration of Table 13.6*b* in Table 13.8*a*.

TABLE 13.8*a*

INVERSE MATRIX WITHOUT REDUCTION OF I—SQUARE ROOT METHOD

1.000	0.313	0.280	0.495
*	1.000	0.652	0.650
*	*	1.000	0.803
*	*	*	1.000
1.000	0.313	0.280	0.495
	0.950	0.594	0.521
		0.754	0.471
			0.512
1.000			
−0.330	1.053		
−0.112	−0.829	1.326	
−0.529	−0.308	−1.220	1.953
1.401	*	*	*
−0.092	1.891	*	*
0.497	−0.723	3.247	*
−1.033	−0.602	−2.383	3.814

Duncan and Kenney have suggested [D] that a more compact form of the method of Table 13.8*a* is possible. It makes the calculation of s^{-1} easy by recording it in mesh with s. The synthetic solution they suggest then appears as

(2)

	a
a^{-1}	
	s
$(s')^{-1}$	

with the values of the symmetric matrix a recorded above the diagonal and the symmetric matrix a^{-1} recorded below the diagonal. This use of the diagonal and the part above or below the diagonal is not uncommon, particularly if the matrix is of higher order. The solution of Table 13.8*a* is presented in Table 13.8*b*.

Another method that is good for symmetric matrices and does not require the formal reduction of I is based on the properties of the

TABLE 13.8*b*

DUNCAN-KENNEY FORM—INVERSE MATRIX WITH SQUARE ROOT METHOD

1.000	0.313	0.280	0.495
1.401	*	*	*
*	1.000	0.652	0.650
−0.092	1.891	*	*
*	*	1.000	0.803
0.497	−0.723	3.247	*
*	*	*	1.000
−1.033	−0.602	−2.383	3.814
1.000	0.313	0.280	0.495
1.000			
	0.950	0.594	0.521
−0.330	1.053		
		0.754	0.471
−0.112	−0.829	1.326	
			0.512
−0.529	−0.308	−1.220	1.953

methods of section 13.6. If the matrix is of order n, there are $\frac{n(n+1)}{2}$ different elements. Now consider the square root method of Table 13.6*b*. There are $n = 4$ elements in the diagonal that are the reciprocals of the diagonals on the left. In addition there are $\frac{n(n-1)}{2} = \frac{4(3)}{2} = 6$ zero elements above the diagonal. This means that there are $\frac{n(n+1)}{2} = \frac{4(5)}{2} = 10 = 6 + 4$ equations, and these are all that are necessary to determine the ten different a_{ij}. The other $\frac{n(n-1)}{2}$ c_{ij} can be filled in from symmetry. The method advised for Table 13.6*b* did not use the elements below the diagonal of the right-hand matrix at all (although they could be used for checking). Hence we can dispense with the whole right-hand side. It is probably wise to compute the reciprocals of the s_{11}, $s_{22 \cdot 1}$, $s_{33 \cdot 12}$, etc., record them either above the s_{11}, $s_{22 \cdot 1}$, $s_{33 \cdot 12}$ (or in a special column at the right) and to remember that this value must be used in the calculation of any diagonal term. The back solution then proceeds as in Table 13.6*b*. The technique is applied to the problem of Table 13.6*b* in Table 13.8*c*.

TABLE 13.8c

INVERSE WITHOUT REDUCTION OF I—SQUARE ROOT METHOD

1.000	0.313	0.280	0.495	
*	1.000	0.652	0.650	
*	*	1.000	0.803	
*	*	*	1.000	
1.000	0.313	0.280	0.495	(1.000)
	0.950	0.594	0.521	(1.058)
		0.754	0.471	(1.326)
			0.512	(1.953)
1.401	−0.091	0.497	−1.033	
−0.091	1.892	−0.724	−0.602	
0.497	−0.725	3.247	−2.382	
−1.033	−0.602	−2.382	3.814	

The inverse of a symmetric matrix with the Gauss-Doolittle method can be computed similarly without the reduction of I. Again there are $\frac{n(n+1)}{2}$ values of zero, and we need only the n diagonal entries. Now the diagonal entries are $\frac{1}{g_{hh\cdot(h-1)}}$, and these can be readily computed from the g matrix. The back solution proceeds, using the g matrix and the symmetric properties of the inverse. The illustration of Table 13.6a is presented with this method in Table 13.8d.

This version of finding the inverse of a square symmetric matrix was given essentially by Waugh [E.2], who made improvements in the method worked out by Horst [E.1]. Dunlap [E.3] attributes a similar method to Cureton.

The compact method of single division can, of course, be used in case the matrix is symmetric. The solution of the problem of Table 13.8d appears in Table 13.8e.

Waugh and Dwyer [F] have exhibited another method for computing the inverse matrix without formal reduction of the identity matrix when the matrix is non-symmetric. This method is more compact than that of Table 13.7b. It features a scheme in which the inverse matrix is in mesh with the elimination entries. A general description of the form and an application to the illustration of Table 13.7b are given in Table 13.8f. There is some advantage in placing the b matrix below the main

TABLE 13.8*d*

INVERSE WITHOUT REDUCTION OF *I*—GAUSS-DOOLITTLE METHOD

1.000	0.4000	0.5000	0.6000	
*	1.0000	0.3000	0.4000	
*	*	1.0000	0.2000	
*	*	*	1.0000	
1.0000	0.4000	0.5000	0.6000	
1.0000	0.4000	0.5000	0.6000	(1.0000)
	0.8400	0.1000	0.1600	
	1.0000	0.1190	0.1905	(1.1905)
		0.7381	−0.1190	
		1.0000	−0.1612	(1.3548)
			0.5903	
			1.0000	(1.6941)
2.0710	−0.1913	−0.7759	1.0109	
−0.1913	1.2842	−0.2185	−0.3552	
−0.7759	−0.2185	1.3988	0.2731	
−1.0109	−0.3552	0.2731	1.6941	

TABLE 13.8*e*

INVERSION WITHOUT REDUCTION OF *I*—COMPACT METHOD OF SINGLE DIVISION

1.0000	0.4000	0.5000	0.6000	
0.4000	1.0000	0.3000	0.4000	
0.5000	0.3000	1.0000	0.2000	
0.6000	0.4000	0.2000	1.0000	
1.0000	0.4000	0.5000	0.6000	(1.0000)
0.4000	0.8400	0.1190	0.1905	(1.1905)
0.5000	0.1000	0.7381	−0.1612	(1.3548)
0.6000	0.1600	−0.1190	0.5903	(1.6941)
2.0710	−0.1913	−0.7759	−1.0109	
−0.1913	1.2842	−0.2185	−0.3552	
−0.7759	−0.2185	1.3988	0.2731	
−1.0109	−0.3552	0.2731	1.6941	

TABLE 13.8*f*

COMPACT COMPUTATION OF THE TRANSPOSE OF THE INVERSE—COMPACT METHOD OF SINGLE DIVISION

General

a_{11}		a_{12}		a_{13}		a_{14}	
a_{21}		a_{22}		a_{23}		a_{24}	
a_{31}		a_{32}		a_{33}		a_{34}	
a_{41}		a_{42}		a_{43}		a_{44}	
a_{11}	c'_{11}	b_{12}	c'_{12}	b_{13}	c'_{13}	b_{14}	c'_{14}
a_{21}	c'_{21}	$g_{22\cdot1}$	c'_{22}	$b_{23\cdot1}$	c'_{23}	$b_{24\cdot1}$	c'_{24}
a_{31}	c'_{31}	$g_{32\cdot1}$	c'_{32}	$g_{33\cdot12}$	c'_{33}	$b_{34\cdot12}$	c'_{34}
a_{41}	c'_{41}	$g_{42\cdot1}$	c'_{42}	$g_{43\cdot12}$	c'_{34}	$g_{44\cdot123}$	c'_{44}

Illustration

26		−10		15		32	
19		45		−14		−8	
−12		16		27		13	
32		29		−35		28	
26.00000	0.02873	−0.38462	−0.00695	0.57692	0.01825	1.23077	−0.00282
19.00000	0.02436	52.30778	0.01239	−0.47720	0.01441	−0.60000	−0.02287
−12.00000	−0.02302	11.38456	0.01572	39.35575	0.00791	0.87916	0.01991
32.00000	−0.01519	41.30784	0.00419	−33.74934	−0.02041	43.07113	0.02322

diagonal and the g matrix above the diagonal. The scheme of the forward solution here is similar to that used earlier in this chapter. The transpose of the inverse appears in the even-numbered columns of the table. The steps in its calculation are explained below.

The technique is based on the fact that

$$ac = ca = I, \tag{3}$$

from which we get

$$ac = I$$

$$a'c' = I. \tag{4}$$

It follows that

$$a(c')' = I$$

$$a'(c') = I. \tag{5}$$

These facts are exhibited synthetically in the form

a_{11}	c'_{11}	a_{12}	c'_{12}	a_{13}	c'_{13}	a_{14}	c'_{14}
a_{21}	c'_{21}	a_{22}	c'_{22}	a_{23}	c'_{23}	a_{24}	c'_{24}
a_{31}	c'_{31}	a_{32}	c'_{32}	a_{33}	c'_{33}	a_{34}	c'_{34}
a_{41}	c'_{41}	a_{42}	c'_{42}	a_{43}	c'_{43}	a_{44}	c'_{44}

where the first equation of (5) indicates row multiplications and the second equation indicates column multiplication. Thus

$$a_{11}c'_{11} + a_{12}c'_{12} + a_{13}c'_{13} + a_{14}c'_{14} = 1$$

$$a_{11}c'_{21} + a_{12}c'_{22} + a_{13}c'_{23} + a_{14}c'_{24} = 0$$

. . . . . .

$$a_{11}c'_{11} + a_{21}c'_{21} + a_{31}c'_{31} + a_{41}c'_{41} = 1$$

$$a_{11}c'_{12} + a_{21}c'_{22} + a_{31}c'_{32} + a_{41}c'_{42} = 0$$

etc.

The usual elimination process by the compact method of single division leads to the two triangular matrices g and b. Considering the unknowns to be the c'_{ij}, we can write equations (5) as

$$gb(c')' = I \tag{6}$$

$$b'g'(c') = I$$

so that

$$b(c')' = g^{-1} \tag{7}$$

$$g'(c') = (b')^{-1}.$$

The first equation indicates a row multiplication of b and c' and the second equation indicates a column multiplication of g and c'. Separate back solutions can be carried out easily since b and g are triangular matrices, if g^{-1} and $(b')^{-1}$ are known.

The back solution in the first case is the usual back solution and in the second case it is a corresponding back solution by columns. The values of g^{-1} and $(b')^{-1}$ are not necessary and they are not computed in Table 13.8*f* since we know that $\frac{n(n-1)}{2}$ values of g^{-1} and an additional

$\frac{n(n-1)}{2}$ value of $(b')^{-1}$ are zero. If we can determine the n values that go with the diagonal terms we shall have a complete solution. The diagonal terms of the $(b')^{-1}$ matrix are the reciprocals of the diagonal term of the g matrix, as is illustrated in Table 13.7*a*. The computational technique of the back solution then proceeds with the calculation of the diagonal terms, the terms in the last row-by-row back solution, the terms in the last column-by-column back solution, the next diagonal term, etc. Thus in the illustration

$$c'_{44} = (0.02322)$$

$$c'_{43} = 0 - (0.87916)(0.02322) = -0.02041$$

$$c'_{42} = 0 - (-0.60000)(0.02322) - (-0.47720)(-0.02041)$$
$$= 0.00419$$

$$c'_{41} = 0 - (1.23077)(0.02322) - (0.57692)(-0.02041)$$
$$- (-0.38462)(0.00419) = -0.01519$$

$$c'_{34} = \frac{0 - (-33.74934)(0.02322)}{39.35575} = 0.01991$$

$$c'_{24} = \frac{0 - (41.30784)(0.02322) - (11.38456)(0.01991)}{52.30778}$$
$$= -0.02267$$

$$c'_{14} = \frac{\left\{0 - (32.00000)(0.02322) - (-12.00000)(0.01991)\right\} - (19.00000)(-0.02267)}{26.00000}$$
$$= -0.00282$$

$$c'_{33} = \frac{1}{39.35575} - (0.87916)(0.01991) = 0.00791,$$

etc.

13.9 The solution of simultaneous equations without a back solution. The methods described in this chapter (which utilize a formal reduction of the identity matrix) can be used at once to eliminate the back solution in solving for the regression coefficient. The square root method is used here to illustrate this technique. It is desired to solve (13.2.1) when a is symmetric. The matrices a, f, and I are placed in

adjacent columns. The square root method is carried on until a is factored. The complete computation then appears as

a	f	I
s	$sa^{-1}f$	sa^{-1}

The premultiplication of $sa^{-1}f$ by $(sa^{-1})'$, which is the equivalent of column-by-column multiplication of sa^{-1} and $sa^{-1}f$, gives the solution

$$a^{-1}s'sa^{-1}f = a^{-1}f = x$$

which is the solution of (13.2.1).

An illustration is presented in Table 13.9a, where the solution of five

TABLE 13.9a

THE ALTERNATIVE TO THE BACK SOLUTION

1.000	0.313	0.280	0.182	0.166	0.495	1	0	0	0	0
*	1.000	0.652	0.554	0.615	0.650	0	1	0	0	0
*	*	1.000	0.747	0.693	0.803	0	0	1	0	0
*	*	*	1.000	0.774	0.804	0	0	0	1	0
*	*	*	*	1.000	0.812	0	0	0	0	1
1.000	0.313	0.280	0.182	0.166	0.495	1.000	0	0	0	0
	0.950	0.594	0.523	0.593	0.521	−0.329	1.053	0	0	0
		0.754	0.511	0.390	0.471	−0.112	−0.830	1.326	0	0
			0.657	0.357	0.306	0.072	−0.193	−1.031	1.522	0
				0.584	0.219	0.081	−0.397	−0.255	−0.930	1.712
0.311	0.012	0.253	0.262	0.375						

equations in five unknowns is shown on the bottom line. Thus

$$b_{16\cdot 2345} = (1.000)(0.495) + (-0.329)(0.521) + (-0.112)(0.471) + (0.072)(0.306) + (0.081)(0.219) = 0.311$$

whereas

$$b_{26\cdot 1345} = (0)(0.495) + (1.053)(0.521) + (-0.830)(0.471) + (-0.193)(0.306) + (-0.397)(0.219) = 0.012,$$

etc.

REFERENCES

A. A. C. Aitken, *Determinants and Matrices*, Oliver and Boyd, London, 1942. See pp. 51–54.

B. P. S. Dwyer, "A matrix presentation of least squares and correlation theory with matrix justification of improved methods of solution," *Annals of Mathematical Statistics*, **15**, 82–89 (1944). See pp. 85–86.

C. P. S. Dwyer, "Recent developments in correlation technique," *Journal of the American Statistical Association*, **37**, 441–460 (1942). See p. 453.

D. D. B. Duncan and J. F. Kenney, *On the Solution of the Normal Equations and Related Topics*, Edwards Brothers, Ann Arbor, 1946. See pp. 28–29.

E. 1. Paul Horst, "A general method for calculating multiple regression constants," *Journal of the American Statistical Association*, **27**, 270–278 (1932).
2. F. V. Waugh, "A simplified method of determining multiple regression constants," *Journal of the American Statistical Association*, **30**, 694–700 (1935).
3. J. W. Dunlap, *Workbook in Statistical Method*, Prentice-Hall, New York, 1939.

F. F. V. Waugh and P. S. Dwyer, "Compact computation of the inverse of a matrix," *Annals of Mathematical Statistics*, **16**, 259–271 (1945). See section 5.

EXERCISES

1. Obtain the adjugate and the inverse of the matrix $\begin{bmatrix} 3 & 1 & 2 \\ 4 & -1 & 3 \\ 2 & 6 & -1 \end{bmatrix}$, and thence obtain the solution of the equations

$$\begin{aligned} 3x + y + 2z &= 12 \\ 4x - y + 3z &= 8 \\ 2x + 6y - z &= 19. \end{aligned}$$

2. Calculate the inverse of the triangular matrix

$$\begin{bmatrix} 1.000 & 0.313 & 0.280 \\ 0 & 0.950 & 0.594 \\ 0 & 0 & 0.754 \end{bmatrix}$$

by the method of Table 13.4*a*. Verify the result.

3. Calculate the inverse of the matrix

$$\begin{bmatrix} 1.0 & 0.4 & 0.5 \\ 0.4 & 1.0 & 0.3 \\ 0.5 & 0.3 & 1.0 \end{bmatrix}$$

by the method of Table 13.6*a*. Verify the result.

4. Calculate the inverse of

$$\begin{bmatrix} 1.000 & 0.313 & 0.280 \\ 0.313 & 1.000 & 0.652 \\ 0.280 & 0.652 & 1.000 \end{bmatrix}$$

by the method of Table 13.6*b*. Verify the result.

5. Calculate the inverse of

$$\begin{bmatrix} 26 & -10 & 15 \\ 19 & 45 & -14 \\ -12 & 16 & 27 \end{bmatrix}$$

by the method of Table 13.6*c*. Verify the result.

6. Calculate the inverse of

$$\begin{bmatrix} 16 & -19 & 40 & -17 \\ 82 & 42 & -32 & -8 \\ 55 & -35 & 45 & -25 \\ 7 & 19 & 33 & 85 \end{bmatrix}$$

by the method of Table 13.6*c*. Verify the result.

7. Calculate the inverse of the problems of exercise 6 by the method of Table 13.7*a*. Verify the result.

8. Calculate the inverse of the problem of exercise 6 by the method of Table 13.7*b*. Verify the result.

9. Calculate the inverse of the matrix of the coefficients of the equations of exercise 6.11 by the method of Table 13.8*a*. Verify the results.

10. Calculate the inverse of the matrix of the coefficients of the equations of exercise 6.11 by the method of Table 13.8*c*. Verify the results.

11. Calculate the inverse of the matrix of the coefficients of the equations of exercise 6.11 by the method of Table 13.8*d*. Verify the results.

12. Calculate the inverse of the matrix of the coefficients of the equations of exercise 5.5 by the method of Table 13.8*e*. Verify the results.

13. Calculate the inverse of the matrix of the coefficients of the equations of exercise 5.5 by the method of Table 13.8*f*. Verify the results.

14. Use the method of Table 13.9*a* and the forward solution of Table 13.9*a* to obtain the solution involving x_1, x_2, x_3, and x_4 if the x_5 values are omitted.

CHAPTER 14

The Calculation of the Adjoint and Inverse with the Method of Determinants

14.1 Introduction. The methods of calculation of the inverse in Chapter 13 are approximate division methods. This chapter presents an outline of similar techniques based on the exact method of determinants. Of course the element of the inverse, being the ratio of two determinants, can be expressed exactly as the ratio of two digital numbers, but the elements of the adjoint are digital numbers. Thus the precise methods of getting the inverse can be based on exact methods of getting the adjoint and dividing by the determinant of the matrix (to any desired number of places) if the matrix is non-singular. If the matrix is singular, the adjoint exists even though the inverse does not. Some of the material in this chapter is presented in less detail than that of Chapter 13, since there is some parallelism as to content.

14.2 The determination of the adjoint with the formal solution of $ax = \Delta I$. The solution of $ax = \Delta I$ is very similar to that of $ax = I$, with the exception that x is now the adjoint A rather than inverse c. Since the adjoint is exact, it is wise to have an exact method available. The method of determinants is the method used as the basis of this presentation.

Some of the illustrations of this chapter are fairly trivial, that is, the coefficients are small positive and negative digits. These illustrations are chosen so that the reader may carry out the calculations of the illustrations without extensive side calculation. Less trivial problems may be carried out, of course, with the use of the proper mechanical equipment.

Two such illustrations are indicated in Tables 14.2*a* and 14.2*b*. The first illustration, which is not symmetric, is taken from Burington [A]. The second illustration is symmetric. The formal solution in each case proceeds with the compact method of determinants. The back solution of the first column on the right leads to the first row of A', the second column to the second row of A', etc. The process should terminate

TABLE 14.2*a*

CALCULATION OF THE ADJOINT—METHOD OF DETERMINANTS

2	1	−2	8	0	0
1	1	1	0	8	0
−1	−2	3	0	0	8
2	1	−2	8	0	0
1	1	4	−8	16	0
−1	−3	8	−8	24	8
5	−4	−1			
1	4	3			
3	−4	1			

TABLE 14.2*b*

CALCULATION OF THE ADJOINT—SYMMETRIC METHOD OF DETERMINANTS

4	2	1	36	0	0
2	4	2	0	36	0
1	2	4	0	0	36
4	2	1	36	0	0
2	12	6	−72	144	0
1	6	36	0	−216	432
12	−6	0			
−6	15	−6			
0	−6	12			

with a final verification that shows that the rows of a, multiplied by the rows of A', give ΔI. Symmetry may be used in Table 14.2*b* in getting A'. The adjoint is the transpose of A'.

Some computers may prefer to use a less compact form of the method of determinants, particularly if the order is high. This can be done, and the first row and column of each matrix can be inserted in the form above before the back solution is started.

14.3 The double-bordering method. A method that is mechanically easy for the non-symmetric case, but is not very satisfactory because

of the amount of space and recording demanded, is the double-bordering method. The square matrix is bordered below as well as on the right with the identity matrix, and a zero matrix completes the square. The method of determinants is then carried through the number of steps necessary to reduce the original matrix, and the matrix appearing in the lower right-hand corner is then $-A$. An illustration to the problem of Table 14.2*a* is given in Table 14.3*a*. The general form is comparable to

TABLE 14.3*a*

THE DOUBLE-BORDERING METHOD

2	1	−2	1	0	0
1	1	1	0	1	0
−1	−2	3	0	0	1
1	0	0	0	0	0
0	1	0	0	0	0
0	0	1	0	0	0
2	1	−2	1	0	0
1	1	4	−1	2	0
−1	−3	8	−1	3	1
1	−1	3	−5	−1	−3
0	2	−4	4	−4	4
0	0	1	1	−3	−1

that of Table 13.7*b*, with the exception that the method of determinants, rather than the method of single division, is used. Perhaps the simplest way to justify this scheme is to note that each of the n^2 determinants in the right lower corner is the negative of the cofactor of the corresponding element. Thus

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} & 1 \\ a_{21} & a_{22} & a_{23} & 0 \\ a_{31} & a_{32} & a_{33} & 0 \\ 1 & 0 & 0 & 0 \end{vmatrix} = - \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} = - A_{11}$$

and

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} & 0 \\ a_{21} & a_{22} & a_{23} & 1 \\ a_{31} & a_{32} & a_{33} & 0 \\ 1 & 0 & 0 & 0 \end{vmatrix} = \begin{vmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{vmatrix} = - A_{12}.$$

The last calculational terms may be added instead of subtracting them from zero, to get the A rather than the $-A$.

14.4 The determination of the adjoint without a back solution when a is symmetric. The solution of the last section is much more satisfactory if a is symmetric. In this case it is not necessary to reduce the second identity matrix since its reduction is the transpose of that of the first matrix. The calculational form features the typical forward solution of a symmetrix matrix and I with the method of determinants, followed by typical calculations with signs changed. The method is illustrated in Table 14.4a.

TABLE 14.4a

ABBREVIATED DOUBLE-BORDERING METHOD FOR SYMMETRIC MATRICES

4	2	1	1	0	0
2	4	2	0	1	0
1	2	4	0	0	1
4	2	1	1	0	0
2	12	6	−2	4	0
1	6	36	0	−6	12
			12	−6	0
			*	15	−6
			*	*	12

The detailed calculation of the last matrix on the right is given here.

$$\frac{(1)^2(12)+(-2)^2}{4}=4, \qquad \frac{(4)(36)+(0)^2}{12}=12,$$

$$\frac{(1)(0)(12)+(-2)(4)}{4}=-2, \qquad \frac{(-2)(36)+(0)(-6)}{12}=-6,$$

$$\frac{(1)(0)(12)+(-2)(0)}{4}=0, \qquad \frac{(0)(36)+(0)(12)}{12}=0,$$

$$\frac{(0)(12)+(4)^2}{4}=4, \qquad \frac{(4)(36)+(-6)^2}{12}=15,$$

$$\frac{(0)(0)(12)+(4)(0)}{4}=0, \qquad \frac{(0)(36)+(-6)(12)}{12}=-6,$$

$$\frac{(0)(12)+(0)^2}{4}=0, \qquad \frac{(0)(36)+(12)^2}{12}=12.$$

The application of this method to a less trivial problem is shown in Table 14.4*b*. The result is the adjoint of a, and the inverse of a may be

TABLE 14.4*b*

SECOND ILLUSTRATION OF ABBREVIATED DOUBLE-BORDERING METHOD

1.0	0.4	0.5	0.6	1	0	0	0
*	1.0	0.3	0.4	0	1	0	0
*	*	1.0	0.2	0	0	1	0
*	*	*	1.0	0	0	0	1
1.0	0.4	0.5	0.6	1	0	0	0
	0.84	0.10	0.16	−0.40	1	0	0
		0.62	−0.10	−0.38	−0.10	0.84	0
			0.366	−0.370	−0.130	0.100	−0.620
				0.758	−0.070	−0.284	−0.370
				*	0.470	−0.080	−0.130
				*	*	0.512	0.100
				*	*	*	0.620

computed correctly to any desired number of places by dividing the elements of A by $\Delta = 0.366$.

Another method that avoids a back solution if the matrix is symmetric is illustrated in Table 14.4*c*. The forward solution is identical with that of Table 14.4*a*, but there are some additional entries. The entries in the middle matrix on the right are multiplied by the value of the determinant, and certain entries are placed at the left of the row. The

TABLE 14.4*c*

SECOND METHOD OF CALCULATING ADJOINT OF SYMMETRIC MATRICES

	4	2	1	1		0		0
	2	4	2	0		1		0
	1	2	4	0		0		1
4	4	2	1	1	36	0	0	0	0
48	2	12	6	−2	−72	4	144	0	0
432	1	6	36	0	0	−6	−216	12	432
				12		−6		0	
				*		15		−6	
				*		*		12	

first of these is the diagonal term of the first row, and the others are the products of the diagonal term of the row with the diagonal term of the row above.

In this case we find the element in row i and column j of A (or A') by multiplying the first elements in column i of the second matrix on the right by the second elements of column j, dividing by the entry at the left, and summing. Thus

$$\frac{(1)(36)}{4} + \frac{(-2)(-72)}{48} + \frac{(0)(0)}{432} = 12$$

$$\frac{(1)(0)}{4} + \frac{(-2)(144)}{48} + \frac{(0)(-216)}{432} = -6$$

$$\frac{(0)(0)}{4} + \frac{(4)(144)}{48} + \frac{(-6)(-216)}{432} = 15.$$

These calculations can be carried out as operational units, U_{14}, if the machine is equipped with a device for disconnecting the revolutions register while each product is being formed.

TABLE 14.4*d*

CALCULATION OF ADJOINT AND INVERSE BY SECOND METHOD

	1.0	0.4	0.5	0.6	1	0	0	0
	*	1.0	0.3	0.4	0	1	0	0
	*	*	1.0	0.2	0	0	1	0
	*	*	*	1.0	0	0	0	1
1.0	1.0	0.4	0.5	0.6	1	0	0	0
0.84		0.84	0.10	0.16	−0.40	1.0	0	0
0.5208			0.62	−0.10	−0.38	−0.10	0.84	0
0.22692				0.366	−0.370	−0.130	0.100	0.620
	2.0710	−0.1913	−0.7760	−1.0109	0.758	−0.070	−0.284	−0.370
	−0.1913	1.2842	−0.2186	−0.3552	*	0.470	−0.080	−0.130
	−0.7760	−0.2186	1.3989	0.2732	*	*	0.512	0.100
	−1.0109	−0.3552	0.2732	1.6940	*	*	*	0.620

This is an exact method even though all the divisions are not necessarily exact as is illustrated in Table 14.4*d* where some of the divisions are approximate. However, the terms in the adjoint have a fixed number of digits (in this case three since the elements are one-digit numbers)

so that it is necessary only to carry the division to four or five places and round off the sum to the three-digit value. The value of $c = c'$ (to four decimals) as well as the value of $A = A'$ is given in Table 14.4*d*.

We justify this method with the use of matrices by relating it to the square root method. We know that

$$s_{ij} = \frac{d_{ij}}{\sqrt{d_{ii}}}$$

$$s_{ij\cdot(h)} = \frac{d_{ij\cdot(h)}}{\sqrt{d_{ii\cdot(h)}d_{hh\cdot(h-1)}}},$$

where $d_{ij} = a_{ij}$. This can be written as

(1) $$d = Ds \quad \text{and} \quad s = D^{-1}d,$$

where D is the diagonal matrix $\begin{bmatrix} \sqrt{d_{11}} & 0 & 0 & \cdots \\ 0 & \sqrt{d_{11}d_{22\cdot 1}} & 0 & \cdots \\ 0 & 0 & \sqrt{d_{22\cdot 1}d_{33\cdot 12}} & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}$.

Now the method of determinants reduces the basic equation $ax = I$ to

(2) $$dx = da^{-1} = Dsa^{-1}$$

so that the matrices in mesh by columns on the right side are Dsa^{-1} and $Dsa\Delta$. If each of these matrices is premultiplied by D^{-1}, we have sa^{-1} and $sa^{-1}\Delta$ so that a column-by-column multiplication gives $a^{-1}s'sa^{-1}\Delta = a^{-1}\Delta = A$. This is the equivalent of the foregoing technique, but, instead of dividing each term separately by the proper term of D (which involves square roots and demands additional recording), we first form the terms of the columnar products and divide each by the proper term of the diagonal matrix $D'D = D^2$.

14.5 The calculation of the adjoint matrix without the reduction of the identity matrix. We may omit the reduction of the identity matrix with the method of determinants just as with the method of single division. This is more easily done if the matrix is symmetric, but it can also be done with non-symmetric matrices. No attempt is made here to exhaust the various possibilities, but a recommended compact tech-

nique is made for symmetric matrices, and a somewhat more general technique is made for non-symmetric matrices. The reader is referred to the solution of Table 14.4*a*. In the general case the forward solution takes the form of Table 14.5*a*.

TABLE 14.5*a*

ADJOINT MATRIX (SYMMETRIC) WITH REDUCTION OF ΔI

a_{11}	a_{12}	a_{13}	a_{14}	Δ	0	0	0
*	a_{22}	a_{23}	a_{24}	0	Δ	0	0
*	*	a_{33}	a_{34}	0	0	Δ	0
*	*	*	a_{44}	0	0	0	Δ
d_{11}	d_{12}	d_{13}	d_{14}	Δ	0	0	0
	$d_{22\cdot1}$	$d_{23\cdot1}$	$d_{24\cdot1}$	x	$d_{11}\Delta$	0	0
		$d_{33\cdot12}$	$d_{34\cdot12}$	x	x	$d_{22\cdot1}\Delta$	0
			$d_{44\cdot123}$	x	x	x	$d_{33\cdot12}\Delta$

In this case there are $\dfrac{n(n-1)}{2}$ zero terms and n diagonal terms. The solution proceeds very much as the solution of Table 13.8*c*. The values of Δ, $d_{11}\Delta$, $d_{22\cdot1}\Delta$ can be placed in an additional column or they can be placed above the diagonal term. The solution of the problem of Table 14.4*d* appears in Table 14.5*b*.

TABLE 14.5*b*

ILLUSTRATION

1.0	0.4	0.5	0.6	
*	1.0	0.3	0.4	
*	*	1.0	0.2	
*	*	*	1.0	
1.0	0.4	0.5	0.6	0.366
	0.84	0.10	0.16	0.366
		0.62	−0.10	0.30744
			0.366	0.22692
0.758	−0.070	−0.284	−0.370	
−0.070	0.470	−0.080	−0.130	
−0.284	−0.080	0.512	0.100	
−0.370	−0.130	0.100	0.620	

$$A_{44} = \frac{0.22692}{0.366} = 0.620$$

$$A_{34} = A_{43} = \frac{0 - (-0.10)(0.620)}{0.62} = 0.100$$

$$A_{24} = A_{42} = \frac{0 - (0.16)(0.620) - (0.10)(0.100)}{0.84} = -0.130$$

$$A_{14} = A_{41} = \frac{0 - (0.6)(0.620) - (0.5)(0.100) - (0.4)(-0.130)}{1.0}$$

$$= -0.370$$

$$A_{33} = \frac{0.30744 - (-0.10)(0.100)}{0.62} = 0.512.$$

The fact that all these divisions must be exact is used as a check.

An alternative form in which the d matrix and the resultant A matrix are in mesh by columns is indicated in Table 14.5c. The details of the calculation are identical with those of Table 14.5b.

TABLE 14.5c

ALTERNATIVE FORM FOR ILLUSTRATION OF TABLE 14.5b

1.0		0.4		0.5		0.6		
*		1.0		0.3		0.4		
*		*		1.0		0.2		
*		*		*		1.0		
1.0000	0.758	0.400	−0.070	0.500	−0.284	0.600	−0.370	0.366
	−0.070	0.840	0.470	0.100	−0.080	0.160	−0.130	0.366
	−0.284		−0.080	0.620	0.512	−0.100	0.100	0.30744
	−0.370		−0.130		0.100	0.366	0.620	0.22692

A generalization of the method of Tables 14.5a, 14.5b, and 14.5c has been made for non-symmetric matrices, using the method of determinants. It has been described by Waugh and Dwyer [B]. This method is similar in spirit to the method of section 13.8, but it uses the method of determinants rather than the method of single division. The detailed steps of the calculation are indicated, in the general case, when $n = 4$, in Table 14.5d.

TABLE 14.5*d*

ADJOINT WITH METHOD OF DETERMINANTS—WAUGH AND DWYER

General							
a_{11}		a_{12}		a_{13}		a_{14}	
a_{21}		a_{22}		a_{23}		a_{24}	
a_{31}		a_{32}		a_{33}		a_{34}	
a_{41}		a_{42}		a_{43}		a_{44}	
d_{11}	A'_{11}	d_{12}	A'_{12}	d_{13}	A'_{13}	d_{14}	A'_{14}
d_{21}	A'_{21}	$d_{22\cdot1}$	A'_{22}	$d_{23\cdot1}$	A'_{23}	$d_{24\cdot1}$	A'_{24}
d_{31}	A'_{31}	$d_{32\cdot1}$	A'_{32}	$d_{33\cdot12}$	A'_{33}	$d_{34\cdot12}$	A'_{34}
d_{41}	A'_{41}	$d_{42\cdot1}$	A'_{42}	$d_{43\cdot12}$	A'_{43}	$d_{44\cdot123}$	A'_{44}

The forward solution elimination process is carried out with the method of determinants although columns are left for the insertion of A' as in Table 13.8*f* and in Table 14.5*c*. The values of Δ, $d_{11}\Delta$, $d_{22\cdot1}\Delta$, $d_{33\cdot12}\Delta$ may be placed on the right as in Table 14.5*c*, and the bottom row is calculated as in Table 14.5*c*. The last column is then computed by a back solution, using columns as in Table 13.8*f*.

$$A'_{44} = \frac{d_{33\cdot12}\Delta}{d_{44\cdot123}} = d_{33\cdot12}$$

$$A'_{43} = \frac{0 - d_{34\cdot12}A'_{44}}{d_{33\cdot12}} = -d_{34\cdot12}$$

$$A'_{42} = \frac{0 - d_{24\cdot1}A'_{44} - d_{23\cdot1}A'_{43}}{d_{22\cdot1}} = d_{24\cdot13}$$

$$A'_{41} = \frac{0 - d_{14}A'_{44} - d_{13}A'_{43} - d_{12}A'_{42}}{d_{11}} = -d_{14\cdot23}$$

$$A'_{34} = \frac{0 - d_{43\cdot12}A'_{44}}{d_{33\cdot12}} = -d_{43\cdot12}$$

$$A'_{24} = \frac{0 - d_{42\cdot1}A'_{44} - d_{32\cdot1}A'_{34}}{d_{22\cdot1}} = d_{42\cdot13}$$

$$A'_{14} = \frac{0 - d_{41}A'_{44} - d_{31}A'_{34} - d_{21}A'_{24}}{d_{11}} = -d_{41\cdot23}$$

$$A'_{33} = \left[\begin{array}{c} \dfrac{d_{22\cdot 1}\Delta - A'_{34}d_{34\cdot 12}}{d_{33\cdot 12}} \\ \dfrac{d_{22\cdot 1}\Delta - A'_{43}d_{43\cdot 12}}{d_{33\cdot 12}} \end{array}\right] = d_{44\cdot 12},$$

etc.

The method is illustrated in Table 14.5*e*. We note that the result as recorded is the adjugate or the transpose of the adjoint and that the inverse can be computed to any desired accuracy by transposing rows and columns and dividing by Δ. The fact that all the operations involve exact divisions can be used as a continuous check. It is also to be

TABLE 14.5*e*

ILLUSTRATION OF TABLE 14.5*d*

26		−10		15		32	
19		45		−14		−8	
−12		16		27		13	
32		29		−35		28	
26	66233	−10	−16033	15	42069	32	−6503
19	56151	1360	28558	−649	33194	−816	−52258
−12	−53068	296	36236	53524	18224	47056	45899
32	−35013	1074	9659	−45899	−47056	2305327	53524

noted that the values of $A'_{44} = d_{33\cdot 12}$, $A'_{34} = -d_{34\cdot 12}$, $A'_{43} = -d_{43\cdot 12}$ were present in the forward solution and that $A'_{33} = d_{44\cdot 12}$ was used in computing $d_{44\cdot 123}$.

Some computers may prefer some less compact form of the method of determinants. The forward solution may be worked out in more detail. The forward solution above may then be prepared from the first row and column of each matrix in the more detailed solution.

The mathematical justification is similar to that of the general method of section 13.8. The equations (13.8.5) become

$$\begin{aligned} a(A')' &= \Delta I \\ a'(A') &= \Delta I \end{aligned} \tag{1}$$

so that

$$\begin{aligned} d(A')' &= da^{-1}\Delta \\ d'(A') &= d'(a^{-1})'\Delta. \end{aligned} \tag{2}$$

The right side of each equation of (2) is a triangular matrix. The back solution of the first of these equations is used to get the rows, and the back solution of the second of these equations is used to get the columns.

REFERENCES

A. R. S. Burington, *Handbook of Mathematical Tables and Formulas*, Handbook Publishers, Inc., Sandusky, Ohio, Third Edition, 1948. See p. 5.

B. F. V. Waugh and P. S. Dwyer, "Compact computation of the inverse of a matrix," *Annals of Mathematical Statistics*, **16**, 259–271 (1945). See section 7.

EXERCISES

1. Calculate the adjoint of the determinant of the coefficients of the equations of Table 4.10*a* by the method of Table 14.2*a*. Verify the results.

2. Calculate the adjoint

$$\begin{bmatrix} 1.0 & 0.4 & 0.5 \\ 0.4 & 1.0 & 0.3 \\ 0.5 & 0.3 & 1.0 \end{bmatrix}$$

by the method of Table 14.5*c*. Verify the results.

3. Calculate the adjoint of the problem of exercise 2 by the method of Table 14.4*c*. Verify the results.

4. Calculate the adjoint of the problem of exercise 2 by the method of Table 14.5*a*. Verify the results.

5. Calculate the adjoint of the problem of exercise 2 by the method of Table 14.5*d*. Verify the results.

6. Calculate the adjoint of the matrix of the coefficients of the equations of exercise 4.10 by the method of Table 14.5*d*. Verify the results.

CHAPTER 15

Problems Involving the Characteristic Equation

15.1 Introduction. This chapter is devoted to the numerical solution of linear problems of the general form

$$
\begin{aligned}
a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + a_{14}x_4 &= \lambda x_1 \\
a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + a_{24}x_4 &= \lambda x_2 \\
a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + a_{34}x_4 &= \lambda x_3 \\
a_{41}x_1 + a_{42}x_2 + a_{43}x_3 + a_{44}x_4 &= \lambda x_4
\end{aligned} \tag{1}
$$

which appear frequently in mathematical statistics and in other fields of applied mathematics. Homogeneous equations of this type can be written in the form

$$
\begin{aligned}
(a_{11} - \lambda)x_1 + a_{12}x_2 + a_{13}x_3 + a_{14}x_4 &= 0 \\
a_{21}x_1 + (a_{22} - \lambda)x_2 + a_{23}x_3 + a_{24}x_4 &= 0 \\
a_{31}x_1 + a_{32}x_2 + (a_{33} - \lambda)x_3 + a_{34}x_4 &= 0 \\
a_{41}x_1 + a_{42}x_2 + a_{43}x_3 + (a_{44} - \lambda)x_4 &= 0.
\end{aligned} \tag{2}
$$

The trivial solution of $x_1 = x_2 = x_3 = x_4 = 0$ always exists. It is desirable to find if there are other solutions. The condition for this is that the determinant of the coefficients

$$
\begin{vmatrix}
a_{11} - \lambda & a_{12} & a_{13} & a_{14} \\
a_{21} & a_{22} - \lambda & a_{23} & a_{24} \\
a_{31} & a_{32} & a_{33} - \lambda & a_{34} \\
a_{41} & a_{42} & a_{43} & a_{44} - \lambda
\end{vmatrix} = 0. \tag{3}
$$

In general the value of this determinant is a polynomial in λ, and the condition (3) results in a series of values, $\lambda_1, \lambda_2, \lambda_3, \cdots$, that satisfy (3).

Each of these values can, in turn, be placed in (2), and the different sets of homogeneous equations can be solved by usual methods.

Direct computational methods which are useful in connection with this problem are presented in this chapter. These methods feature the calculation of the adjoint (and inverse) of a as well as the solutions of its characteristic equation.

15.2 The characteristic equation. The *characteristic equation* arises from the writing of (15.1.3) in polynomial form. Thus the characteristic equation can be written, somewhat symbolically, as

$$\lambda^4 - \{1\}\lambda^3 + \{2\}\lambda^2 - \{3\}\lambda^3 + \{4\} = 0 \tag{1}$$

where $\{1\}$, $\{2\}$, $\{3\}$, and $\{4\}$ are functions of the coefficients a_{ij}.

In the more general case (15.1.1) may be written in matrix form as

$$ax = \lambda Ix \tag{2}$$

where a is a square n by n matrix, x is a one by n matrix (column vector), and I is a diagonal matrix with each diagonal term equal to λ. In a similar manner (15.1.2) becomes

$$(a - \lambda I)x = 0. \tag{3}$$

The condition (15.1.3) is

$$|a - \lambda I| = 0. \tag{4}$$

The expanded form of (4) is commonly known as the characteristic equation.

An alternative form for (4) is

$$|\lambda I - a| = 0. \tag{5}$$

These two characteristic equations are equivalent since each term of the left side of (5) is the negative of the corresponding term of the left side of (4). The characteristic equation appears either as

$$|a - \lambda I| = \{n\} - \lambda\{n-1\} + \lambda^2\{n-2\} + \cdots + (-1)^k\lambda^k\{n-k\} + \cdots + (-1)^{n-1}\lambda^{n-1}\{1\} + (-1)^n\lambda^n = 0 \tag{6}$$

or as

$$|\lambda I - a| = \lambda^n - \lambda^{n-1}\{1\} + \lambda^{n-2}\{2\} + \cdots + (-1)^k\lambda^{n-k}\{k\} + \cdots + (-1)^{n-1}\lambda\{n-1\} + (-1)^n\{n\} = 0. \tag{7}$$

These forms have been used by writers whose methods are discussed in this chapter.

15.3 The coefficients of the characteristic equation as the sums of principal minors. The coefficients in (15.2.6) and (15.2.7), aside from sign, are the sums of all principal minors * of a having an order indicated by the symbol $\{\ \}$. Thus the coefficient of λ^k in (6) is, aside from the sign, the sum of all the principal minors of a having order $n - k$, and the coefficient of λ^{n-k} in (7) is, aside from the sign, the sum of all the principal minors of order k. The value of $\{n\}$ is thus the value of the determinant; the value of $\{n-1\}$ is the sum of all the minors of the elements of the principal diagonal. The type of proof is illustrated in (1) and (2), where $n = 3$.

$$\begin{vmatrix} a_{11} - \lambda & a_{12} + 0 & a_{13} + 0 \\ a_{21} + 0 & a_{22} - \lambda & a_{23} + 0 \\ a_{31} + 0 & a_{32} + 0 & a_{33} - \lambda \end{vmatrix} = 0. \tag{1}$$

The determinant on the left can be written as the sum of $2^3 = 8$ determinants, each taking on an a or λ column from the three positions. Expansion shows that

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} - \lambda \left\{ \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} + \begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix} + \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \right\} + \lambda^2(a_{33} + a_{22} + a_{11}) - \lambda^3 = 0. \tag{2}$$

The method of section 10.10 may be used to obtain all the principal minors that may be added to obtain the absolute values of the coefficients of the characteristic equation. Thus the characteristic equation of Table 10.10*b* is (to three decimal places)

$$\lambda^4 - 1.307\lambda^3 + 0.428\lambda^2 - 0.026\lambda + 0.000 = 0. \tag{3}$$

A further illustration is given in Table 15.3*a*, where the method of determinants is applied to a problem used by Hotelling to illustrate another method. The sums of all the principal minors of a given order are indicated in the last column so that it is easy to present the characteristic equation in the form of (15.2.7) by using alternating signs. Thus the equation in Table 15.3*a* is

$$\lambda^5 - 5\lambda^4 + 33\lambda^3 - 51\lambda^2 + 135\lambda + 225 = 0. \tag{4}$$

In factored form this appears as

$$(\lambda + 1)(\lambda^2 - 3\lambda + 15)^2 = 0. \tag{5}$$

* A *principal minor* of a determinant is a minor whose principal diagonal elements are on the principal diagonal of the determinant.

TABLE 15.3*a*

CHARACTERISTIC EQUATION—METHOD OF DETERMINANTS

HOTELLING ILLUSTRATION

15	11	6	−9	−15	
1	3	9	−3	−8	
7	6	6	−3	−11	5
7	7	5	−3	−11	
17	12	5	−10	−16	
(15)	34	129	−36	−105	
	13	48	18	−60	
	28	33	18	−60	
	−7	−27	3	15	
(3)		−36	9	15	
		−48	12	23	33
		−93	6	48	
(6)			−3	−11	
			−45	−41	
(−3)				−62	
(34)		−3	72	−45	
		−166	108	60	
		−1	−10	−15	
(48)			18	−60	
			42	−60	
(18)				30	51
(−36)			0	−36	
			207	−111	
(12)				146	
(−3)				−62	
(−3)			342	−225	
			3	0	
(108)				−30	135
(18)				30	
(0)				−207	
(342)				−225	−225

Other direct elimination methods could be used in evaluating the principal minor, but the method of determinants seems to be the desirable direct method as it uses the results of determinants of lower order in evaluating the determinants of higher order and has the useful checking device of exact division.

15.4 The Bingham method for obtaining the adjoint and characteristic equation. Hotelling, Bingham, and Girshick [A.1, 2] have worked out a method for obtaining the inverse of a matrix based on the characteristic equation and on the Cayley-Hamilton theorem that states that every matrix satisfies its own characteristic equation. They write

$$(1)\quad |a - \lambda I| = e_n - e_{n-1}\lambda + e_{n-2}\lambda^2 + \cdots + (-1)^k e_{n-k}\lambda^k + \cdots + (-1)^{n-1}e_1\lambda^{n-1} + (-1)^n\lambda^n = 0$$

where the e_i are the usual elementary symmetric functions of the roots of the characteristic equation and e_n is the determinant of a. The form of (1) is similar to that of (15.2.7). They then use the Cayley-Hamilton theorem to get

$$(2)\quad e_n I - e_{n-1}a + e_{n-2}a^2 + \cdots + (-1)^k e_{n-k}a^k + \cdots + (-1)^{n-1}e_1 a^{n-1} + (-1)^n a^n = 0.$$

A postmultiplication by a^{-1}, if a^{-1} exists, and rearrangement of the terms result in

$$(3)\quad A = \Delta a^{-1} = e_n a^{-1} = e_{n-1}I - e_{n-2}a + \cdots + (-1)^{k-2}e_{n-k}a^{k-1} + \cdots + (-1)^{n-1}e_1 a^{n-2} + (-1)^n a^{n-1}$$

which shows that the adjoint (and inverse) is a linear function of the values of e_k and the powers of a to $n - 1$.

The authors advise the use of Newton's identities [A],

$$(4)\quad \begin{aligned} e_1 &= S_1 \\ e_2 &= \tfrac{1}{2}(e_1 S_1 - S_2) \\ e_3 &= \tfrac{1}{3}(e_2 S_1 - e_1 S_2 + S_3) \\ e_4 &= \tfrac{1}{4}(e_3 S_1 - e_2 S_2 + e_1 S_3 - S_4) \\ &\cdots\cdots\cdots\cdots \\ e_k &= \frac{1}{k}(e_{k-1}S_1 - e_{k-2}S_2 + \cdots \pm S_k). \end{aligned}$$

They calculate the values a^2, a^3, $\cdots$, a^{n-1} and obtain the values of S_r with the formula

$$S_r = \text{trace } a^r, \tag{5}$$

where the trace of a^r is the sum of the terms in the principal diagonal of a^r.

TABLE 15.4*a*

HOTELLING-BINGHAM-GIBSHICK METHOD

						S_i	c_i	
a	15	11	6	−9	−15			
	1	3	9	−3	−8			
	7	6	6	−3	−11			−51
	7	7	5	−3	−11			
	17	12	5	−10	−16	5	5	
a^2	−40	−9	105	−9	−40			
	−76	−43	32	44	23			
	−55	−22	62	20	−10			33
	−61	−25	65	20	−7			
	−40	−9	110	−14	−40	−41	33	
a^3	−617	−380	64	499	256			
	−260	−189	−316	355	280			
	−443	−279	−106	415	259			−5
	−464	−300	−136	439	292			
	−617	−385	69	499	256	−217	51	
a^4	−1342	−978	−2963	2444	2006			
	944	522	−1982	−10	503			
	−358	−333	−2435	1307	1334			1
	−175	−243	−2645	1247	1355			
	−1312	−963	−2978	2444	1991	−17	135	
a^5	9361							
		6024						
			−6550					
				−7052				
					1402	3185	−225	
A	−207	64	−124	111	171			
	−315	30	195	−180	270			
	−315	30	−30	45	270			
	−225	75	−75	0	225			
	−414	53	52	−3	342			

A computing form for this method as applied to the Hotelling illustration is given in Table 15.4a. The matrix is multiplied by itself to get the matrix a^2. This is done by writing each row of a' on a slip of paper and placing it directly under each row of a, in turn, so that a row-by-row multiplication of a' and a can be made. This is continued through the calculation of a^2, a^3, and a^4 and the diagonal terms of a^5. The values of S_i are found and then the value of e_i. The characteristic equation is

$$\lambda^5 - 5\lambda^4 + 33\lambda^3 - 51\lambda^2 + 135\lambda + 225 = 0 \tag{6}$$

as indicated in the last section.

The value of A is then found by using

$$A = e_4 I - e_3 a + e_2 a^2 - e_1 a^3 + a^4 = 135I - 51a + 33a^2 - 5a^3 + a^4.$$

The coefficients to be used as multipliers are indicated in the last column, and the value of the adjoint is recorded in the last rows.

15.5 The Frame method of obtaining the adjoint and characteristic equation. Frame [B] has recently developed a recursion formula for obtaining the adjoint, determinant, inverse, and characteristic equation of a matrix that serves as the basis of an excellent computational technique. His form of the characteristic equation is

$$|\lambda I - a| = F(\lambda) = \lambda^n - c_1\lambda^{n-1} - c_2\lambda^{n-2} - \cdots - c_k\lambda^{n-k} - \cdots - c_{n-1}\lambda - c_n = 0. \tag{1}$$

Comparison with (15.2.7) shows that

$$c_k = (-1)^{k-1}\{k\}. \tag{2}$$

Now the minor of any element of the matrix $(\lambda I - a)$ cannot have any power of λ greater than $n - 1$. Frame then writes

$$\text{adj } (\lambda I - a) = C(\lambda) = \lambda^{n-1}a_0 + \lambda^{n-2}a_1 + \cdots + \lambda^{n-k}a_k + \cdots + \lambda a_{n-2} + a_{n-1} \tag{3}$$

where the a_k are square matrices that serve as undetermined multipliers. It is desired to determine these values of a_k and also the values of c_k.

The following proof holds for non-singular matrices. We make use of the fact (13.2.5) that a matrix postmultiplied by its adjoint is equal to the determinant of the matrix times the identity matrix and get

$$(\lambda I - a) \text{ adj } (\lambda I - a) = (\lambda I - a)C(\lambda) = |\lambda I - a|\, I = F(\lambda)I. \tag{4}$$

Substituting (1) and (3) in (4) and equating coefficients of λ^0, λ, λ^2, and λ^k, we get

$$
\begin{aligned}
a_0 &= I \\
a_1 &= aa_0 - c_1 I \\
a_2 &= aa_1 - c_2 I \\
&\cdots\cdots\cdots \\
a_k &= aa_{k-1} - c_k I \\
&\cdots\cdots\cdots \\
0 &= aa_{n-1} - c_n I.
\end{aligned} \tag{5}
$$

Continued substitution using (5) shows that

$$a_k = a^k - c_1 a^{k-1} - c_2 a^{k-2} - \cdots - c_{k-1} a - c_k I \tag{6}$$

for $k = 1, 2, \cdots, n-1$. The values of a_k are then computable, if we know the value of c_k, since the powers of a are computable. The determination of the c's, however, permits the use of the recursion formula (5) as the basis of a computing technique.

Now it can be shown that the trace of $C(\lambda)$ is equal to the derivative of $F(\lambda)$:

$$F'(\lambda) = \text{trace } C(\lambda). \tag{7}$$

From (4) we know

$$C(\lambda) - aC(\lambda) = F(\lambda) I.$$

Taking the trace of each side, we get

$$\text{trace } C(\lambda) - \text{trace } aC(\lambda) = \text{trace } F(\lambda) I$$

so that

$$\lambda F'(\lambda) - \text{trace } aC(\lambda) = nF(\lambda)$$

and

$$
\begin{aligned}
\text{trace } aC(\lambda) &= \lambda F'(\lambda) - nF(\lambda) \\
&= \Sigma\,[-(n-k)c_k \lambda^{n-k} + nc_k \lambda^{n-k}] \\
&= \Sigma k c_k \lambda^{n-k}.
\end{aligned} \tag{8}
$$

Using the value of $C(\lambda)$ and equating coefficients of λ^{n-k}, we get

$$\text{trace } aa_{k-1} = kc_k \quad \text{and} \quad c_k = \text{trace } \frac{aa_{k-1}}{k}. \tag{9}$$

Equations (5) and (9) enable us to calculate the values of a_i and c_i.

starting with a and $a_0 = I$. Note from (1), with $\lambda = 0$, that

$$|-a| = -c_n.$$

An equivalent statement is obtained from (2), where

$$c_n = (-1)^{n-1}\{n\}, \tag{10}$$

so that

$$c_n = (-1)^{n-1}\Delta \qquad \text{and} \qquad \Delta = (-1)^{n-1}c_n.$$

Similarly, $\lambda = 0$ in (3) results in

$$adj(-a) = a_{n-1}, \quad \text{so that} \quad A = adj(a) = (-1)^{n-1}a_{n-1}. \tag{11}$$

The inverse is, of course, A/Δ.

The recursion process features the determination of the adjoint, the determinant, and the inverse as well as the characteristic equation.

The following computing form is based on Frame's recursion formulas.

(*a*) Write the matrix a and follow it with a column that gives the diagonal terms of a. Add the diagonal terms to get aa_0 and enter this sum in the row beneath it. Divide by one to get c_1 and place this result in the next column.

(*b*) Write a', the transpose of a, directly beneath a. Decrease the diagonal terms of a by c_1. Call this a'_1. Multiply the corresponding rows of a and a'_1 to get the diagonal terms of aa_1, which are now recorded in the column at the right of a'. Add these to get the trace of aa_1, and divide by two to get the value of c_2. Record these values at the right of the row below a'. The values of the sums of the columns of a'_1 may be used as the basis of a column sum check if desired.

The diagonal values of a'_2 are then found by subtracting c_2 from the terms in the column to the right of a'. The other terms in a'_2 are obtained by a row-by-row multiplication of a_1 and a'_1. A corresponding calculation is made for the column sums, and the column sum check is applied. The value of c_2 should be subtracted in getting c'. This process is continued until a_{n-1} is reached. It can be carried one step further and can be used as a check since $a'_n = 0$.

An illustration to a simple problem proposed by Frame is given in

Table 15.5*a*. The characteristic equation is

$$\lambda^4 - 5\lambda^3 + 6\lambda^2 + 4\lambda - 8 = 0. \tag{12}$$

The value of A is the negative of the transpose of a'_3. The value of the determinant is 8.

A corresponding treatment of a non-symmetric illustration used in previous chapters is given in Table 15.5*b*. It is, of course, not necessary

TABLE 15.5*a*

MODIFIED FRAME METHOD WITH COLUMN SUM CHECKS—FRAME ILLUSTRATION

a	1	−2	3	−2	1	
	1	5	−1	−1	5	
	2	3	2	−2	2	
	2	−2	6	−3	−3	
					5	5
a'_1	−4	1	2	2	−4	
	−2	0	3	−2	−3	
	3	−1	−3	6	−15	
	−2	−1	−2	−8	10	
	−5	−1	0	−2	−12	−6
a'_2	2	−3	−5	−4	1	
	11	3	6	20	0	
	−16	−5	−9	−28	−9	
	10	3	5	16	−4	
	7	−2	−3	4	−12	−4
a'_3	5	−4	−7	−8	8	
	−17	4	3	−8	8	
	23	−4	−5	8	8	
	−13	4	7	0	8	
	−2	0	−2	−8	32	8
a'_4	0	0	0	0	0	
	0	0	0	0	0	
	0	0	0	0	0	
	0	0	0	0	0	
	0	0	0	0	0	0

to record the 0 values of the a'_4. The characteristic equation is

$$\lambda^4 - 126\lambda^3 + 6088\lambda^2 - 166539\lambda + 2{,}305{,}327 = 0. \tag{13}$$

A slightly modified presentation, which does not exhibit the column sum checks, is given in Table 15.5*c*, where application is made to the illustration of Tables 15.3*a* and 15.4*a*. The values of c_i are placed in additional columns on the right, thus minimizing the number of rows necessary for the solution.

TABLE 15.5*b*

MODIFIED FRAME METHOD—WAUGH-DWYER ILLUSTRATION

a	26	−10	15	32	26	
	19	45	−14	−8	45	
	−12	16	27	13	27	
	32	29	−35	28	28	
					126	126
a'_1	−100	19	−12	32	−1946	
	−10	−81	16	29	−4291	
	15	−14	−99	−35	−3532	
	32	−8	13	−98	−2407	
	−63	−84	−82	−72	−12176	−6088
a'_2	4142	−1133	1596	−1333	100306	
	1718	1797	−367	−2417	137981	
	−2075	1321	2556	2559	148315	
	−2029	850	−1435	3681	113015	
	1756	2835	2350	2490	499617	166539
a'_3	−66233	16033	−42069	6503	−2305327	
	−56151	−28558	−33194	52258	−2305327	
	53068	−36236	−18224	−45899	−2305327	
	35013	−9659	47056	−53524	−2305327	
	−34303	−58420	−46431	−40662	−9221308	−2305327
a'_4	0	0	0	0		
	0	0	0	0		
	0	0	0	0		
	0	0	0	0		
	0	0	0	0		

TABLE 15.5c

MODIFIED FRAME METHOD—HOTELLING ILLUSTRATION

	15	11	6	−9	−15	15		
	1	3	9	−3	−8	3		
a	7	6	6	−3	−11	6		
	7	7	5	−3	−11	−3		
	17	12	5	−10	−16	−16	5	5
	10	1	7	7	17	−115		
	11	−2	6	7	12	−58		
a'_1	6	9	1	5	5	32		
	−9	−3	−3	−8	−10	35		
	−15	−8	−11	−11	−21	40	−66	−33
	−82	−81	−90	−96	−125	78		
	−64	−25	−52	−60	−69	125		
a'_2	75	−13	65	40	85	−218		
	36	59	35	68	36	240		
	35	63	45	48	73	−72	153	51
	27	153	63	72	144	−342		
	28	74	29	56	56	−105		
a'_3	−263	−179	−269	−269	−316	−165		
	247	36	216	189	239	−135		
	−39	−99	−54	−36	−123	207	−540	−135
	−207	−315	−315	−225	−414	−225		
	64	30	30	75	53	−225		
a'_4	−124	195	−30	−75	52	−225		
	111	−180	45	0	−3	−225		
	171	270	270	225	342	−225	−1125	−225
	0	0	0	0	0	0		
	0	0	0	0	0	0		
a'_5	0	0	0	0	0	0		
	0	0	0	0	0	0		
	0	0	0	0	0	0	0	0

15.6 The characteristic vectors. The problem proposed in (15.2.3) involves more than the determination of the characteristic equation. This equation may be solved by the usual synthetic methods. Each of the values of λ_i is then placed in (15.2.3), and the resulting set of homogeneous equations is solved. The resulting solution for a single λ_i is known as a *characteristic vector* or a *modal column*.

The conventional solutions differ appreciably for different types of solutions of the characteristic equation. A discussion of this situation is not given in detail here, but is available in excellent form in a book by Frazer, Duncan, and Collar [C]. However, the Frame method is readily adapted to this problem, and a few illustrations show how simply this can be done. One of the chief advantages of the Frame method is the fact that calculations leading to the characteristic equation are themselves used in getting the characteristic vectors.

The basic theorem needed, applicable to the simplest and commonest case, is that the characteristic vectors for given λ_i are proportional to the columns of adj $(\lambda I - a)$. Now this enables us to apply (15.5.3) immediately to our computational form. We have

$$(1) \quad \text{adj } (\lambda I - a) = \lambda^{n-1} I + \lambda^{n-2} a_1 + \lambda^{n-3} a_2 + \cdots + \lambda^{n-k-1} a_k + \cdots + a_{n-1}.$$

The value of a_i may be multiplied by the appropriate power of λ and the results added.

This computing form uses the values of a'_i so the rows of a'_i rather than the columns of a may be used in obtaining the characteristic vectors.

The method is illustrated in Table 15.6*a*, where the method is applied to an illustration of Frazer, Duncan, and Collar [C]. The method of the previous section provides a characteristic equation with the roots 1, -2, and 3. The values of $I + a'_1 + a'_2$, $4I - 2a'_1 + a'_2$, $9I + 3a'_1 + a'_2$ are indicated in columns on the right. The results are indicated in the bottom rows. Thus the characteristic vectors are indicated by $(1, -1, -1)$, $(11, 1, -14)$, $(1, 1, \frac{1}{2})$. It is not necessary to calculate all the nine rows.

Two further illustrations are given in Tables 15.6*b* and 15.6*c*. In this case the characteristic equation has a multiple root, in which an illustration of Frazer, Duncan, and Collar [C] is used.

Fettis also developed substantially the same method [D].

TABLE 15.6*a*

CHARACTERISTIC VECTORS—FRAME METHOD (FIRST ILLUSTRATION)

a	$\begin{matrix}2 & -2 & 3\\ 1 & 1 & 1\\ 1 & 3 & -1\end{matrix}$	$\begin{matrix}2\\1\\-1\end{matrix}$		I	$4I$	$9I$
		2	2			
a'_1	$\begin{matrix}0 & 1 & 1\\ -2 & -1 & 3\\ 3 & 1 & -3\end{matrix}$	$\begin{matrix}1\\0\\9\end{matrix}$		1	-2	3
	$1 \quad 1 \quad 1$	10	5			
a'_2	$\begin{matrix}-4 & 2 & 2\\ 7 & -5 & -8\\ -5 & 1 & 4\end{matrix}$	$\begin{matrix}-6\\-6\\-6\end{matrix}$		1	1	1
	$-2 \quad -2 \quad -2$	-18	-6			
	$\begin{matrix}0 & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & 0\end{matrix}$	$\begin{matrix}0\\0\\0\end{matrix}$				
	$\lambda^3 - 2\lambda^2 - 5\lambda + 6 = 0, \quad \lambda = 1$				-2	3
	$\begin{matrix}-3 & 3 & 3\\ 5 & -5 & -5\\ -2 & 2 & 2\end{matrix}$					
	$\begin{matrix}0 & 0 & 0\\ 11 & 1 & -14\\ -11 & -1 & 14\end{matrix}$					
	$\begin{matrix}5 & 5 & 5\\ 1 & 1 & 1\\ 4 & 4 & 4\end{matrix}$					

TABLE 15.6*b*

CHARACTERISTIC VECTORS WITH FRAME METHOD—FRAZER-DUNCAN-COLLAR
(SECOND ILLUSTRATION)

a	2	−2	3	2		I	$4I$
	10	−4	5	−4			
	5	−4	6	6			
				4	4		
a'_1	−2	10	5	−9		1	2
	−2	−8	−4	−8			
	3	5	2	7			
	−1	7	3	−10	−5		
a'_2	−4	−35	−20	2		1	1
	0	−3	−2	2			
	2	20	12	2			
	−2	−18	−10	6	2		
	0	0	0				
	0	0	0				
	0	0	0				

$\lambda^3 - 4\lambda^2 + 5\lambda - 2 = 0, \quad \lambda = 1, 1, 2$

−5	−25	−15	
−2	−10	−6	$\lambda = 1$
5	25	15	
−4	−15	−10	
−4	−15	−10	$\lambda = 2$
8	30	20	

TABLE 15.6*c*

CHARACTERISTIC VECTOR (THIRD ILLUSTRATION)

a	1	−2	3	−2	1			
	1	5	−1	−1	5			
	2	3	2	−2	2		$-I$	$8I$
	2	−2	6	−3	−3			
					5	5		
a'_1	−4	1	2	2	−4			
	−2	0	3	−2	−3			
	3	−1	−3	6	−15		1	4
	−2	−1	−2	−8	10			
	−5	−1	0	−2	−12	−6		
a'_2	2	−3	−5	−4	1			
	11	3	6	20	0			
	−16	−5	−9	−28	−9		−1	2
	10	3	5	16	−4			
	7	−2	−3	4	−12	−4		
a'_3	5	−4	−7	−8	8			
	−17	4	3	−8	8			
	23	−4	−5	8	8		I	1
	−13	4	7	0	8			
	−2	0	−2	−8	32	8		
a'_4	0	0	0	0				
	0	0	0	0				
	0	0	0	0				
	0	0	0	0				
	0	0	0	0		0		

$\lambda^4 - 5\lambda^3 + 6\lambda^2 + 4\lambda - 8 = 0$

$\lambda = -1, 2, 2, 2$

	−2	0	0	−2	
	−30	0	0	−30	
	42	0	0	42	$\lambda = -1$
	−25	0	0	−25	
	1	−6	−9	−8	
	−3	18	27	24	
	3	−18	−27	−24	$\lambda = 2$
	−1	6	9	8	

REFERENCES

A. 1. M. D. Bingham, "A new method for obtaining the inverse matrix," *Journal of the American Statistical Association*, **36**, 530–534 (1941).

2. H. Hotelling, "Some new methods in matrix calculation," *Annals of Mathematical Statistics*, **14**, 1–34 (1943). See section 11.

B. J. S. Frame, *A Simple Recursion Formula for Inverting a Matrix.* Presented to American Mathematical Society at Boulder, Colorado, on Sept. 1, 1949.

C. R. A. Frazer, W. J. Duncan, and A. R. Collar, *Elementary Matrices*, Cambridge University Press, 1947. See pp. 64–82.

D. H. E. Fettis, "A method for obtaining the characteristic equation of a matrix and computing the associated modal columns," *Quarterly of Applied Mathematics*, **8**, 206–212 (1950).

EXERCISES

Find the characteristic equation by the method of determinants of:

1. The matrix of the coefficients in Table 4.2*a*.
2. The matrix of the coefficients in Table 4.3*a*.
3. The matrix of the coefficients in Table 4.8*a*.
4. The matrix of the coefficients in Table 4.12*b*.
5. The matrix of the coefficients in Table 4.13*a*.
6. The matrix of the coefficients in exercise 4.1.
7. The matrix of the coefficients in exercise 4.10.

Find the characteristic equation by the method of Table 15.4*a*:

8. For the problem of exercise 1.
9. For the problem of exercise 2.
10. For the problem of exercise 3.
11. For the problem of exercise 4.
12. For the problem of exercise 5.
13. For the problem of exercise 6.
14. For the problem of exercise 7.

Find the characteristic equation by the modified Frame method:

15. For the problem of exercise 1.
16. For the problem of exercise 2.
17. For the problem of exercise 3.
18. For the problem of exercise 4.
19. For the problem of exercise 5.
20. For the problem of exercise 6.
21. For the problem of exercise 7.

Obtain characteristic vectors, with the Frame method:

22. For the problem of exercise 2.
23. For the problem of exercise 3.
24. For the problem of exercise 4.
25. For the problem of exercise 4.1.
26. For the problem of exercise 4.10.

CHAPTER 16

Other Methods

16.1 Introduction. Many diverse approaches to the problem of solving simultaneous linear equations (and the allied problems of determinants, inverse, and characteristic equation) have been devised, and it is impossible in this book to review all of them. Pivotal elimination methods are extensively described as it is felt that these direct methods are most useful to those who work with desk computing machines. Some authors feel that other types of methods are better, and most agree that various types of mechanical, electrical, or electronic equipment make feasible the use of many methods not recommended for use with the desk calculator. Although this treatment cannot be exhaustive, it does seem appropriate here to present some material on (*a*) extension methods and (*b*) iterative methods, since some authors seem to prefer these even for desk calculators. Extension methods are discussed first since they seem to be more closely related to the direct methods of the earlier chapter.

16.2 Extension methods. Extension methods are enlargement methods. In using these methods, we start with the solution for a single variable (or element of a matrix) and enlarge the solution to include an additional variable (or an additional row and column) of a matrix at each step until the desired calculation is complete. Different terminology has been used to identify this general method. Guttman [A.1] calls it an enlargement method. Frazer, Duncan, and Collar [A.2] call it the method of submatrices, and others, noting the staircase form of some of the results, call it a staircase or escalator method [A.3]. Some authors use exact methods in building the successive terms; others use approximate methods.

We may use the recursion formula (6.3.9) or the identities of section 8.3 as the basis of an enlargement method for the solution of simultaneous equations. A more practical method is based on the technique of section 13.9, although this calls for the reduction of the identity matrix. The solution of the related equations obtained by the omission of x_5

is available from Table 13.9*a*. Thus

$$b_{16\cdot 234} = (1.000)(0.495) + (-0.329)(0.521) + (-0.112)(0.471) + (0.072)(0.306) = 0.293$$

$$b_{26\cdot 134} = (0)(0.495) + (1.053)(0.521) + (-0.830)(0.471) + (-0.193)(0.306) = 0.099,$$

etc.,

so that

$$b_{16\cdot 2345} = b_{16\cdot 234} + (0.081)(0.219)$$

$$b_{26\cdot 1345} = b_{26\cdot 134} + (-0.397)(0.219),$$

etc.

The terms necessary to extend the solution are readily available from Table 13.9*a*.

The calculation of the inverse matrix by extension methods is also advocated. The basic formula needed calls for the inverse of

$$M = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \tag{1}$$

where a, b, c, d are submatrices of the whole matrix and have no reference to the values of a, b, c, d used in earlier chapters. M, of course, is a square non-singular matrix. The results indicate that a and d should also be square. In the common system of enlargement with one variable at a time, c is a row vector, b is a column vector, and d is a scalar. In the common case in which M is symmetric, $c = b'$, and the matrix takes the form

$$M = \begin{bmatrix} a & b \\ b' & d \end{bmatrix}. \tag{2}$$

The problem is easily solved with a little matrix manipulation, if we use the general method of formal reduction of the identity matrix as we did in Chapters 13 and 14. We can write the matrix equations

$$\begin{aligned} ax + by &= Iz + 0w \\ cx + dy &= 0z + Iw \end{aligned} \tag{3}$$

where a and d are square matrices. The solution of (3) can be written in the form

$$\begin{aligned} Ix + 0y &= (\ )z + (\ )w \\ 0x + Iy &= (\ )z + (\ )w \end{aligned} \tag{4}$$

and the matrix coefficients of z and w constitute the inverse matrix of I

We multiply the first equation of (3) by $-ca^{-1}$ and add to the second equation of (3) to get

$$(d - ca^{-1}b)y = -ca^{-1}z + w. \tag{5}$$

Premultiplication by $(d - ca^{-1}b)^{-1}$ gives

$$y = -(d - ca^{-1}b)^{-1}ca^{-1}z + (d - ca^{-1}b)^{-1}w. \tag{6}$$

Substitution of (6) in the first equation of (3) gives

$$ax = Iz + b(d - ca^{-1}b)ca^{-1}z - b(d - ca^{-1}b)^{-1}w \tag{7}$$

so that

$$x = a^{-1}[I + b(d - ca^{-1}b)^{-1}ca^{-1}]z - a^{-1}b(d - ca^{-1}b)^{-1}w. \tag{8}$$

Now (8) and (6) are the equations (4) so that we can write

$$M^{-1} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} \tag{9}$$

$$= \begin{bmatrix} a^{-1}[I + b(d - ca^{-1}b)^{-1}ca^{-1}] & -a^{-1}b(d - ca^{-1}b)^{-1} \\ -(d - ca^{-1}b)^{-1}ca^{-1} & (d - ca^{-1}b)^{-1} \end{bmatrix}.$$

M, a, and $(d - ca^{-1}b)$ must be non-singular, and d must be square. A different order of elimination gives the alternate form

$$M^{-1} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \tag{10}$$

$$= \begin{bmatrix} (a - bd^{-1}c)^{-1} & -(a - bd^{-1}c)bd^{-1} \\ -d^{-1}c(a - bd^{-1}c)^{-1} & a^{-1}[I + c(a - bd^{-1}c)^{-1}b] \end{bmatrix}.$$

The identification of equivalent terms of M^{-1} in (9) and (10) gives identities that can be used in certain computational problems [B].

The usual procedure calls for the use of (9) rather than (10).

If one new variable is added at each step, b and c are vectors so that d, $d - ca^{-1}b$, and $(d - ca^{-1}b)^{-1}$ are scalars. We set $(d - ca^{-1}b)^{-1} = k^2$ and have

$$M^{-1} = \begin{bmatrix} a^{-1} + k^2a^{-1}bca^{-1} & -k^2a^{-1}b \\ -k^2ca^{-1} & k^2 \end{bmatrix}. \tag{11}$$

This is the basis of first-order enlargement and one of the simplest applications of the method of submatrices [C]. Here $a^{-1}b$ is the solu-

tion of the equation $ax = b$, so that (11) can be written

$$M^{-1} = \begin{bmatrix} a^{-1} + k^2 xca^{-1} & -k^2 x \\ -k^2 ca^{-1} & k^2 \end{bmatrix}. \tag{12}$$

If M is symmetric, then $c = b'$ and we can write

$$M^{-1} = \begin{bmatrix} a^{-1} + k^2 xx' & -k^2 x \\ -k^2 x' & k^2 \end{bmatrix} \tag{13}$$

and the answer is written in terms of the inverse of a, the solution of $ax = b$, and $k^2 = (d - b'a^{-1}b)^{-1}$. This can be translated at once to the results of the square root method since the value of k is the reciprocal of the corresponding diagonal entry in the square root method. The values of k^2, $-k^2x'$, and k^2xx' can be substituted in (13) to get the extension.

This method does not translate readily to suitable operational units. A better expansion, using the square root method, is based on the fact that $(s^{-1})(s^{-1})' = a^{-1}$ so that a row-by-row multiplication of s^{-1} gives a^{-1}. We get the last row of each of the successive matrices s_1^{-1}, s_2^{-1}, s_3^{-1}, etc., by back solutions from the forward solution of the square root method. We need only to add to a_i^{-1} the value of $s_{i+1}^{-1}(s_{i+1}^{-1})'$, where s_{i+1}^{-1} is a matrix that has all the elements zero except those of the last rows, which are the elements indicated.

An illustration is presented in Table 16.2*a*. The symmetric matrix M is subjected to the square root process. Back solution gives the successive last rows of s_i^{-1}. These are applied to build up the values of M_2^{-1}, M_3^{-1}, and M_4^{-1}. More explicitly, the value M_4^{-1} is obtained with the use of

$$\begin{bmatrix} 1.121 & -0.255 & -0.149 & 0 \\ -0.255 & 1.796 & -1.099 & 0 \\ -0.149 & -1.099 & 1.758 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & -0.529 \\ 0 & 0 & 0 & -0.308 \\ 0 & 0 & 0 & -1.220 \\ 0 & 0 & 0 & 1.953 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -0.529 & -0.308 & -1.220 & 1.953 \end{bmatrix}.$$

Symmetry is used in presenting the results in Table 16.2*a*.

Exact methods can be used. Duncan, Frazer, and Collar [C] have exhibited an exact method for obtaining the inverse that maintains its

and the matrix coefficients of z and w constitute the inverse matrix of M.

We multiply the first equation of (3) by $-ca^{-1}$ and add to the second equation of (3) to get

$$(d - ca^{-1}b)y = -ca^{-1}z + w. \tag{5}$$

Premultiplication by $(d - ca^{-1}b)^{-1}$ gives

$$y = -(d - ca^{-1}b)^{-1}ca^{-1}z + (d - ca^{-1}b)^{-1}w. \tag{6}$$

Substitution of (6) in the first equation of (3) gives

$$ax = Iz + b(d - ca^{-1}b)ca^{-1}z - b(d - ca^{-1}b)^{-1}w \tag{7}$$

so that

$$x = a^{-1}[I + b(d - ca^{-1}b)^{-1}ca^{-1}]z - a^{-1}b(d - ca^{-1}b)^{-1}w. \tag{8}$$

Now (8) and (6) are the equations (4) so that we can write

$$M^{-1} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} \tag{9}$$

$$= \begin{bmatrix} a^{-1}[I + b(d - ca^{-1}b)^{-1}ca^{-1}] & -a^{-1}b(d - ca^{-1}b)^{-1} \\ -(d - ca^{-1}b)^{-1}ca^{-1} & (d - ca^{-1}b)^{-1} \end{bmatrix}.$$

M, a, and $(d - ca^{-1}b)$ must be non-singular, and d must be square. A different order of elimination gives the alternate form

$$M^{-1} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \tag{10}$$

$$= \begin{bmatrix} (a - bd^{-1}c)^{-1} & -(a - bd^{-1}c)bd^{-1} \\ -d^{-1}c(a - bd^{-1}c)^{-1} & a^{-1}[I + c(a - bd^{-1}c)^{-1}b] \end{bmatrix}.$$

The identification of equivalent terms of M^{-1} in (9) and (10) gives identities that can be used in certain computational problems [B].

The usual procedure calls for the use of (9) rather than (10).

If one new variable is added at each step, b and c are vectors so that d, $d - ca^{-1}b$, and $(d - ca^{-1}b)^{-1}$ are scalars. We set $(d - ca^{-1}b)^{-1} = k^2$ and have

$$M^{-1} = \begin{bmatrix} a^{-1} + k^2a^{-1}bca^{-1} & -k^2a^{-1}b \\ -k^2ca^{-1} & k^2 \end{bmatrix}. \tag{11}$$

This is the basis of first-order enlargement and one of the simplest applications of the method of submatrices [C]. Here $a^{-1}b$ is the solu-

tion of the equation $ax = b$, so that (11) can be written

$$M^{-1} = \begin{bmatrix} a^{-1} + k^2xca^{-1} & -k^2x \\ -k^2ca^{-1} & k^2 \end{bmatrix}. \tag{12}$$

If M is symmetric, then $c = b'$ and we can write

$$M^{-1} = \begin{bmatrix} a^{-1} + k^2xx' & -k^2x \\ -k^2x' & k^2 \end{bmatrix} \tag{13}$$

and the answer is written in terms of the inverse of a, the solution of $ax = b$, and $k^2 = (d - b'a^{-1}b)^{-1}$. This can be translated at once to the results of the square root method since the value of k is the reciprocal of the corresponding diagonal entry in the square root method. The values of k^2, $-k^2x'$, and k^2xx' can be substituted in (13) to get the extension.

This method does not translate readily to suitable operational units. A better expansion, using the square root method, is based on the fact that $(s^{-1})(s^{-1})' = a^{-1}$ so that a row-by-row multiplication of s^{-1} gives a^{-1}. We get the last row of each of the successive matrices s_1^{-1}, s_2^{-1}, s_3^{-1}, etc., by back solutions from the forward solution of the square root method. We need only to add to a_i^{-1} the value of $s_{i+1}^{-1}(s_{i+1}^{-1})'$, where s_{i+1}^{-1} is a matrix that has all the elements zero except those of the last rows, which are the elements indicated.

An illustration is presented in Table 16.2*a*. The symmetric matrix M is subjected to the square root process. Back solution gives the successive last rows of s_i^{-1}. These are applied to build up the values of M_2^{-1}, M_3^{-1}, and M_4^{-1}. More explicitly, the value M_4^{-1} is obtained with the use of

$$\begin{bmatrix} 1.121 & -0.255 & -0.149 & 0 \\ -0.255 & 1.796 & -1.099 & 0 \\ -0.149 & -1.099 & 1.758 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & -0.529 \\ 0 & 0 & 0 & -0.308 \\ 0 & 0 & 0 & -1.220 \\ 0 & 0 & 0 & 1.953 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -0.529 & -0.308 & -1.220 & 1.953 \end{bmatrix}.$$

Symmetry is used in presenting the results in Table 16.2*a*.

Exact methods can be used. Duncan, Frazer, and Collar [C] have exhibited an exact method for obtaining the inverse that maintains its

exactness through the use of fractions. Fractions may be avoided by using the adjoint or adjugate. The inverse may be obtained in fractional form, or in decimal form if the approximate operation of division by the determinant is permitted.

TABLE 16.2*a*

EXTENSION USING SQUARE ROOT METHOD

1.000	0.313	0.280	0.495	
*	1.000	0.652	0.650	
*	*	1.000	0.803	M
*	*	*	1.000	
1.000	0.313	0.280	0.495	
	0.950	0.594	0.521	s
		0.754	0.471	
			0.512	
1.000				
−0.330	1.053			Last row
−0.112	−0.829	1.326		of s_i^{-1}
−0.529	−0.308	−1.220	1.953	
1.000				M_1^{-1}
1.109	*			
−0.347	1.109			M_2^{-1}
1.121	*	*		
−0.255	1.796	*		M_3^{-1}
−0.149	−1.099	1.758		
1.401	*	*	*	
−0.092	1.891	*	*	
0.497	−0.723	3.247	*	M_4^{-1}
−1.033	−0.602	−2.383	3.814	

We may build our enlargement process on some forward solution. If the matrix is not symmetric, we build our process on some such exact forward solution as that of Table 14.5*e*. This forward solution is presented in the odd-numbered columns of Table 16.2*b*. Back solutions providing the last row and last column for each adjugate matrix A'_i are then found. We use the adjugate matrix, which is the transpose of the adjoint. Thus $A'_3 = A'_{ij(3)}$ is the transpose of the adjoint of

the matrix

$$a_3 = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}.$$

The elements of A'_3 are obtained by replacing the elements of a by their cofactors. In the notation of the method of determinants, the values of A'_3 become

$$(14) \quad A'_3 = \begin{bmatrix} A'_{11(3)} = d_{22\cdot 3} & A'_{12(3)} = -d_{21\cdot 3} & A'_{13(3)} = d_{31\cdot 2} \\ A'_{21(3)} = -d_{12\cdot 3} & A'_{22(3)} = d_{11\cdot 3} & A'_{23(3)} = -d_{32\cdot 1} \\ A'_{31(3)} = d_{13\cdot 2} & A'_{32(3)} = -d_{23\cdot 1} & A'_{33(3)} = d_{11\cdot 2} \end{bmatrix}.$$

Now the value of $d_{11\cdot 2} = d_{22\cdot 1} = d_{\cdot 12}$ is available in the forward solution. (The values of $A'_{11(1)}$ and the four values of $A'_{ij(2)}$ may be obtained by inspection.) The various results of such operations are indicated in the even-numbered columns of the first part of Table 16.2*b*.

Once these preliminary computations are completed, the actual enlargement process begins. A computing form is prepared in the lower

TABLE 16.2*b*

ENLARGEMENT METHOD FOR ADJUGATE MATRIX

a_{11} a_{21} a_{31} a_{41}		a_{12} a_{22} a_{32} a_{42}		a_{13} a_{23} a_{33} a_{43}		a_{14} a_{24} a_{34} a_{44}	
a_{11} a_{21} a_{31} a_{41}	$A'_{11(1)} = 1$ $A'_{21(2)} = -a_{12}$ $A'_{31(3)} = d_{13\cdot 2}$ $A'_{41(4)} = -d_{14\cdot 23}$	a_{12} $d_{22\cdot 1}$ $d_{32\cdot 1}$ $d_{42\cdot 1}$	$A'_{12(2)} = -a_{21}$ $A'_{22(2)} = a_{11}$ $A'_{32(3)} = -d_{23\cdot 1}$ $A'_{42(4)} = d_{24\cdot 13}$	a_{13} $d_{23\cdot 1}$ $d_{33\cdot 12}$ $d_{43\cdot 12}$	$A'_{13(2)} = d_{31\cdot 2}$ $A'_{23(3)} = -d_{32\cdot 1}$ $A'_{33(3)} = d_{22\cdot 1}$ $A'_{43(4)} = -d_{34\cdot 12}$	a_{14} $d_{24\cdot 1}$ $d_{34\cdot 12}$ $d_{44\cdot 123}$	$A'_{14(4)} = -d_{41\cdot 23}$ $A'_{24(3)} = d_{42\cdot 13}$ $A'_{34(4)} = -d_{43\cdot 12}$ $A'_{44(4)} = d_{33\cdot 12}$
a	$A'_{11(1)} = 1$						
$d_{\cdot 12}$	$A'_{11(2)}$ $A'_{21(2)}$	$A'_{12(2)}$ $A'_{22(2)}$					
$d_{\cdot 123}$	$A'_{11(3)}$ $A'_{21(3)}$ $A'_{31(3)}$	$A'_{12(3)}$ $A'_{22(3)}$ $A'_{32(3)}$	$A'_{13(3)}$ $A'_{23(3)}$ $A'_{33(3)}$				
$d_{\cdot 1234}$	$A'_{11(4)}$ $A'_{21(4)}$ $A'_{31(4)}$ $A'_{41(4)}$	$A'_{12(4)}$ $A'_{22(4)}$ $A'_{32(4)}$ $A'_{42(4)}$	$A'_{13(4)}$ $A'_{23(4)}$ $A'_{33(4)}$ $A'_{43(4)}$	$A'_{14(4)}$ $A'_{24(4)}$ $A'_{34(4)}$ $A'_{44(4)}$			

part of Table 16.2*b* for recording the values of the terms of the adjugates of different order. A column at the left is used for the value of the determinant of each order, and the last row and the last column of each adjugate matrix are filled in from the first part of the table. The other values are then filled in with the use of the formula

$$A'_{ij(h+1)} = \frac{d_{\cdot(h+1)}A'_{ij(h)} + A'_{ih(h+1)}A'_{hj(h+1)}}{d_{\cdot(h)}}, \tag{15}$$

where $d_{\cdot(h+1)}$ is the determinant of order $h + 1$.

The formula (15) leads to a simple computational technique. To find a given term, A'_{ij}, take the corresponding term in the adjugate above, multiply by the value of the determinant at the left, add the product of the term at the bottom of the row with the one at the right of the column, and divide by the determinant above. The division should be exact.

The technique is illustrated in Table 16.2*c*. Application of the

TABLE 16.2*c*

ENLARGEMENT METHOD FOR ADJUGATE MATRIX—ILLUSTRATION

26		−10		15		32	
19		45		−14		−8	
−12		16		27		13	
32		29		−35		28	
26	1	−10	−19	15	844	32	−6503
19	10	1360	26	−649	−296	−816	−52258
−12	−535	296	649	53524	1360	47056	45899
32	−35013	1074	9659	−45899	−47056	2305327	53524
26	1						
1360	45	−19					
1	10	26					
	1439	−345	844				
53524	510	882	−296				
	−535	649	1360				
	66233	−16033	42069	−6503			
	56151	28558	33194	−52258			
2305327	−53068	36236	18224	45899			
	−35013	9659	−47056	53524			

technique shows, for example,

$$A'_{11(2)} = \frac{(1)(1360) + (10)(-19)}{26} = 45$$

$$A'_{11(3)} = \frac{(45)(53524) + (-535)(844)}{1360} = 1439$$

$$A'_{11(4)} = \frac{(1439)(2305327) + (-35013)(-6503)}{53524} = 66233$$

$$A'_{12(4)} = \frac{(510)(2305327) + (-35013)(-52258)}{53524} = 56151$$

$$A'_{13(4)} = \frac{(-535)(2305327) + (-35013)(45899)}{53524} = -53068,$$

etc.

Formula (15) is a special case of Jacobi's theorem, proofs of which are quite accessible [D.1, 2]. We can establish the formula by relating it to the d's in columns and rows i, j, k, after h eliminations have been made. The determinants appear in the form

$$\begin{matrix} d_{ii\cdot(h)} & d_{ij\cdot(h)} & d_{ik\cdot(h)} \\ d_{ji\cdot(h)} & d_{jj\cdot(h)} & d_{jk\cdot(h)} \\ d_{ki\cdot(h)} & d_{kj\cdot(h)} & d_{kk\cdot(h)}. \end{matrix} \tag{16}$$

We use $d_{ii\cdot(h)}$ as a pivot and arrive at the result

$$d_{ii\cdot(h)}d_{\cdot(h)ijk} = d_{jj\cdot(h)i}d_{kk\cdot(h)i} - d_{jk\cdot(h)i}d_{kj\cdot(h)i}. \tag{17}$$

We note that $d_{jj\cdot(h)i} = d_{\cdot(h)ij}$, $d_{kk\cdot(h)i} = d_{ii\cdot(h)k}$, and solve for $d_{ii\cdot(h)k}$ to get

$$d_{ii\cdot(h)k} = \frac{d_{ii\cdot(h)}d_{\cdot(h)ijk} + d_{jk\cdot(h)i}d_{kj\cdot(h)i}}{d_{\cdot(h)ij}}. \tag{18}$$

We now use the fact that $d_{jk\cdot(h)i} = (-1)^{j+k}A'_{jk\cdot(h)}$ and that $d_{ii\cdot(h)} = A'_{jj\cdot(h)}$ to get

$$A'_{jj(hk)} = \frac{A'_{jj(h)}d_{\cdot(h)ijk} + A'_{jk(hi)}A'_{kj(hi)}}{d_{\cdot(h)ij}} \tag{19}$$

which is a general statement for the diagonal terms of (15). Similarly $d_{ij\cdot(h)}$ may be used as a pivot in (16) to get

$$d_{ij\cdot(h)}d_{\cdot(h)ijk} = d_{ij\cdot(h)k}d_{\cdot(h)ij} - d_{ik\cdot(h)j}d_{kj\cdot(h)i} \tag{20}$$

so that we get the general form of (15), where $i \neq j$:

$$A'_{ij(hk)} = \frac{A'_{ij(h)}d_{\cdot(h)ijk} + A'_{ik(hj)}A'_{kj(hi)}}{d_{\cdot(h)ij}}. \tag{21}$$

A less involved method is possible if the matrix is symmetric since the adjugate then equals the adjoint with $A'_{ij(h)} = A'_{ji(h)} = A_{ij(h)} = A_{ji(h)}$. Then the rows of the back solution only need be calculated, and the values of the back solution may be placed directly in the adjugate matrices in the second part of the computing form as they are calculated. The results are presented in the form of Table 16.2*d*. An illustration is shown in Table 16.2*e*. Thus

$$A_{31(4)} = A_{13(4)} = \frac{A_{13(3)}d_{\cdot 1234} + A_{14(4)}A_{34(4)}}{d_{\cdot 123}}$$

and in the illustration

$$A_{31(4)} = \frac{(-0.380)(0.366) + (-0.370)(0.100)}{0.620} = -0.284.$$

TABLE 16.2*d*

ENLARGEMENT METHOD FOR SYMMETRIC MATRIX

	a_{11}	a_{12}	a_{13}	a_{14}
	*	a_{22}	a_{23}	a_{24}
	*	*	a_{33}	a_{34}
	*	*	*	a_{44}
	a_{11}	a_{12}	a_{13}	a_{14}
		$d_{22\cdot 1}$	$d_{23\cdot 1}$	$d_{24\cdot 1}$
			$d_{33\cdot 12}$	$d_{34\cdot 12}$
				$d_{44\cdot 123}$
a_{11}	$1 = A_{11(1)}$			
$d_{\cdot 12}$	$a_{22} = A_{11(2)}$	*		
	$-a_{12} = A_{21(2)}$	$a_{11} = A_{22(2)}$		
$d_{\cdot 123}$	$d_{33\cdot 2} = A_{11(3)}$	*	*	
	$-d_{21\cdot 3} = A_{12(3)}$	$d_{11\cdot 3} = A_{22(3)}$	*	
	$d_{13\cdot 2} = A_{31(3)}$	$-d_{23\cdot 1} = A_{32(3)}$	$d_{22\cdot 1} = A_{33(3)}$	
$d_{\cdot 1234}$	$A_{11(4)}$	*	*	*
	$A_{21(4)}$	$A_{22(4)}$	*	*
	$A_{31(4)}$	$A_{32(4)}$	$A_{33(4)}$	*
	$A_{41(4)}$	$A_{42(4)}$	$A_{43(4)}$	$A_{44(4)}$

TABLE 16.2e

ENLARGEMENT METHOD FOR SYMMETRIC ADJOINT MATRIX—ILLUSTRATION

	1.0	0.4	0.5	0.6
	*	1.0	0.3	0.4
	*	*	1.0	0.2
	*	*	*	1.0
	1.0	0.4	0.5	0.6
		0.84	0.10	0.16
			0.62	−0.10
				0.366
1.0	1			
0.84	1.0	*		
	−0.4	1.0		
0.62	0.91	*	*	
	−0.25	0.75	*	
	−0.38	−0.10	0.84	
0.366	0.758	*	*	*
	−0.070	0.470	*	*
	−0.284	−0.080	0.512	*
	−0.370	−0.130	0.100	0.620

T. Smith [E] has worked out a method for obtaining the successive values of the adjugate by enlargement directly from the matrix itself without the need of a forward solution. He obtained the last row and last column of each adjugate matrix by a process of bordering and then used the method described above to obtain the interior elements.

The values of $A'_{11(2)} = a_{22}$, $A'_{12(2)} = -a_{21}$, $A'_{21(2)} = -a_{12}$, and $A'_{22(2)} = a_{11}$ are obtained by inspection. These are bordered by the third column and the third row of a. From these values we compute the values of $A'_{13(3)}$, $A'_{23(3)}$, $A'_{33(3)} = d_{12}$, $A'_{31(3)}$, and $A'_{32(3)}$. We remember that $A'_{ij(3)}$ is the cofactor of a_{ij} in

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}.$$

Hence

$$A'_{31(3)} = \begin{vmatrix} a_{12} & a_{13} \\ a_{22} & a_{23} \end{vmatrix} = -a_{22}a_{13} + a_{12}a_{23} = -(A'_{11(2)}a_{13} + A'_{21(2)}a_{23}).$$

Similarly,

$$A'_{32(3)} = -(A'_{21(2)}a_{13} + A'_{22(2)}a_{23})$$

$$A'_{13(3)} = -(A'_{11(2)}a_{31} + A'_{12(2)}a_{32})$$

$$A'_{23(3)} = -(A'_{12(2)}a_{31} + A'_{22(2)}a_{32})$$

whereas

$$A'_{33(3)} = d_{\cdot 12} = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = A'_{21\cdot 2}a_{21} + A'_{22\cdot 2}a_{22}.$$

We see then that the row-by-row multiplication and column-by-column multiplication of the elements of $A'_{ij(2)}$ and the values of a_{i3} and a_{3j} give the negatives of the values of $A'_{3j(3)}$ and $A'_{i3(3)}$, whereas the value of $A'_{33(3)}$ is obtained by the multiplication of the second row of $A_{ij(2)}$ by the corresponding row of a.

We calculate the interior elements of the third-order adjugate with the method of Table 16.2*b*. We border the result with the a_{i4} and a_{4j} terms. The values of $A'_{4j(4)}$ and $A'_{i4(4)}$ are found by row-by-row and column-by-column multiplications. These may be justified by formal expansions. Thus

$$A'_{41(4)} = -\begin{vmatrix} a_{12} & a_{13} & a_{14} \\ a_{22} & a_{23} & a_{24} \\ a_{32} & a_{33} & a_{34} \end{vmatrix}$$

$$= -(A'_{11(3)}a_{14} + A'_{12(3)}a_{24} + A'_{13(3)}a_{34}),$$

etc.

The value of $A_{44(4)} = d_{\cdot 123}$ is found by multiplying the third row of A'_3 by the third row of a_3. The internal values are then filled in and the process proceeds.

A general form of the presentation is given in Table 16.2*f* and an illustration in Table 16.2*g*. Here

$$A'_{31(3)} = -[(45)(15) + (10)(-14)] = -535$$

$$A'_{32(3)} = -[(-19)(15) + (26)(-14)] = 649$$

$$A'_{13(3)} = -[(45)(-12) + (-19)(16)] = 844,$$

etc.

TABLE 16.2*f*

ENLARGEMENT OF ADJUGATE BY T. SMITH'S METHOD

a_{11}	a_{12}	a_{13}	a_{14}
a_{21}	a_{22}	a_{23}	a_{24}
a_{31}	a_{32}	a_{33}	a_{34}
a_{41}	a_{42}	a_{43}	a_{44}

a_{11}	$A'_{11(1)} = 1$

$d_{\cdot 12}$	$A'_{11(2)}$	$A'_{12(2)}$	a_{13}
	$A'_{21(2)}$	$A'_{22(2)}$	a_{23}
	a_{31}	a_{32}	a_{33}

$d_{\cdot 123}$	$A'_{11(3)}$	$A'_{12(3)}$	$A'_{13(3)}$	a_{14}
	$A'_{21(3)}$	$A'_{22(3)}$	$A'_{23(3)}$	a_{24}
	$A'_{31(3)}$	$A'_{32(3)}$	$A'_{33(3)}$	a_{34}
	a_{41}	a_{42}	a_{43}	a_{44}

$d_{\cdot 1234}$	$A'_{11(4)}$	$A'_{12(4)}$	$A'_{13(4)}$	$A'_{14(4)}$
	$A'_{21(4)}$	$A'_{22(4)}$	$A'_{23(4)}$	$A'_{24(4)}$
	$A'_{31(4)}$	$A'_{32(4)}$	$A'_{33(4)}$	$A'_{34(4)}$
	$A'_{41(4)}$	$A'_{42(4)}$	$A'_{43(4)}$	$A'_{44(4)}$

A corresponding method for the enlargement of the transpose of the inverse is available if the successive values of $d_{\cdot 12}$, $d_{\cdot 123}$, $d_{\cdot 1234}$, etc., are known. These may be obtained from the forward solution or they may be available from other sources. The technique for the last row and column is based on the fact that the element in the transpose of the

TABLE 16.2*g*

ENLARGEMENT OF ADJUGATE BY T. SMITH'S METHOD—ILLUSTRATION

26	−10	15	32
19	45	−14	−8
−12	16	27	13
32	29	−35	28

26	1

	45	−19	15
1360	10	26	−14
	−12	16	27

	1439	−345	844	32
	510	882	−296	−8
53524	−535	649	1360	13
	32	29	−35	28

	66233	−16033	42069	−6503
	56151	28558	33194	−52258
2305327	−53068	36236	18224	45899
	−35013	9659	−47056	53524

inverse is equal to the corresponding element of the adjugate divided by the determinant. Thus

$$c'_{31(3)} = \frac{A'_{31(3)}}{d_{\cdot 123}} = -(A'_{11(2)}a_{13} + A'_{21(2)}a_{23})$$

so that

$$c'_{31(3)} = -\frac{d_{\cdot 12}}{d_{\cdot 123}}[c'_{11(2)}a_{13} + c'_{21(2)}a_{23}].$$

In general the technique, when $i \neq j$, is like that of the adjugate, but it involves an additional multiplication by the determinantal ratio. When $i = j = 3$, we have

$$c'_{33(3)} = \frac{d_{\cdot 12}}{d_{\cdot 123}}$$

so that the quantity in the lower right corner, which is the ratio of the two determinants, is the multiplier. It is also true that

$$c'_{44(4)} = \frac{d_{\cdot 123}}{d_{\cdot 1234}},$$

etc.

After the last row and column are computed, the internal terms are computed by a formula that may be considered to be a modification of (15). We divide (15) by $d_{\cdot(h+1)}$ to get

$$c'_{ij(h+1)} = c'_{ij(h)} + c'_{i,h+1(h+1)} \frac{A'_{h+1,j(h+1)}}{d_{\cdot(h)}}. \tag{22}$$

Now

$$\frac{A'_{h+1,j(h+1)}}{d_{\cdot(h)}} = \frac{A'_{h+1,j(h+1)}}{d_{\cdot(h+1)}} \cdot \frac{d_{\cdot(h+1)}}{d_{\cdot(h)}} = \frac{c'_{h+1,j(h+1)}}{c'_{h+1,h+1(h+1)}}$$

so that we have

$$c'_{ij(h+1)} = c'_{ij(h)} + \frac{c'_{i,h+1(h+1)} c'_{h+1,j(h+1)}}{c_{h+1,h+1(h+1)}}. \tag{23}$$

We have, written less formally and in better computational form,

$$c'_{ij} = \frac{c'_{ij)h}(c'_{hh} + c'_{ih}c'_{hj}}{c'_{hh}}, \tag{24}$$

where $c'_{ij)h(}$ is the initial term and h is the added variable.

TABLE 16.2*h*

ENLARGEMENT METHOD FOR THE INVERSE

a			
a_{11}	a_{12}	a_{13}	a_{14}
a_{21}	a_{22}	a_{23}	a_{24}
a_{31}	a_{32}	a_{33}	a_{34}
a_{41}	a_{42}	a_{43}	a_{44}

$\Delta_1 = a_{11}$	$c'_{11(1)} = 1$

$d_{.12}$	$c'_{11(2)}$	$c'_{12(2)}$	a_{13}
	$c'_{21(2)}$	$c'_{22(2)}$	a_{23}
	a_{31}	a_{32}	a_{33}

$d_{.123}$	$c'_{11(3)}$	$c'_{12(3)}$	$c'_{13(3)}$	a_{14}
	$c'_{21(3)}$	$c'_{22(3)}$	$c'_{23(3)}$	a_{24}
	$c'_{31(3)}$	$c'_{32(3)}$	$c'_{33(3)}$	a_{34}
	a_{41}	a_{42}	a_{43}	a_{44}

$d_{.1234}$	$c'_{11(4)}$	$c'_{12(4)}$	$c'_{13(4)}$	$c'_{14(4)}$
	$c'_{21(4)}$	$c'_{22(4)}$	$c'_{23(4)}$	$c'_{24(4)}$
	$c'_{31(4)}$	$c'_{32(4)}$	$c'_{33(4)}$	$c'_{34(4)}$
	$c'_{41(4)}$	$c'_{42(4)}$	$c'_{43(4)}$	$c'_{44(4)}$

The computational form is shown in Table 16.2*h* and an illustration in Table 16.2*i*. In this case a is symmetric so that the symmetry property could be used.

Enlargement methods are sometimes called escalator methods since the successive enlargements take on the form of an escalator. A descrip-

TABLE 16.2*i*

ENLARGEMENT METHOD OF THE INVERSE—ILLUSTRATION

1.0	0.4	0.5	0.6
0.4	1.0	0.3	0.4
0.5	0.3	1.0	0.2
0.6	0.4	0.2	1.0

$d_{11} = 1.0$	1.0

$d_{\cdot 12} = 0.84$	1.1905	−0.4762	0.5
	−0.4762	1.1905	0.3
	0.5	0.3	1.0

$d_{\cdot 123} = 0.62$	1.4678	−0.4033	−0.6129	0.6
	−0.4033	1.2097	−0.1613	0.4
	−0.6129	−0.1613	1.3548	0.2
	0.6	0.4	0.2	1.0

$d_{\cdot 1234} = 0.366$	2.0711	−0.1914	−0.7759	−1.0109
	−0.1914	1.2841	−0.2186	−0.3551
	−0.7759	−0.2186	1.3989	0.2732
	−1.0109	−0.3551	0.2732	1.6940

tion of the escalator process as applied to the solution of linear simultaneous equations may be found in *The Escalator Method in Engineering Vibration Problems* [A.3]. An escalator process for the solutions of the characteristic equation, under certain general conditions, is also given there. The reader is also referred to an article by Cochran [F.1] and to work by Boschan [F.2].

16.3 Iterative methods. Iterative methods are also very useful in connection with the basic linear problems of the solution of equations, the calculation of the inverse matrix, and the determination of characteristic vectors. They are especially valuable when the value of p is large, when machines for handling the successive iterations effectively are available, when a good approximation to the answer is known, when we wish to obtain solutions of adjusted equations, or when it is desired to obtain some particular answer and not whole sets of answers.

The author feels that linear problems that should be undertaken with a hand machine are, for the most part, problems for which direct pivotal methods are satisfactory. Some authors, Hotelling, for instance, who has worked out a number of ingenious iterative techniques [G], feel that iterative methods have an important role even in computations with desk machines.

The purpose here is to mention some available iterative techniques suitable to linear problems and to present suitable references.

Iterative methods are excellent for obtaining the approximate roots of numbers, as indicated in Chapter 1. The classical iterative method for solving linear equations that feature symmetry and relatively large diagonal terms is the method of Gauss and Seidel [H]. A first approximation to the solution is obtained by dividing the terms on the right by the corresponding diagonal terms. Then the other $p - 1$ values of x_i are substituted in the first equation to obtain x'_1, this and the other $p - 2$ values of x_i are substituted in the second equation to obtain x'_2, etc. The process is continued until the values of x_i converge to the solution. Hotelling discusses the method with extensive references to proofs of convergence and provides an acceleration of and extension of the method [I].

The Hardy Cross iterative process has been used quite extensively by engineers [J]. This utilizes the successive values of x_1, $x_2 - x_1$, $x_3 - x_2$, $\cdots$ $x_p - x_{p-1}$, etc., instead of the successive values x_1, x_2, x_3, $\cdots$ x_p. It has been generalized by Morris [K].

In general, iterative methods used in solving equations are also useful in determining the inverse. Starting with c_0, an approximation to the inverse of a, Hotelling suggests the calculation of

$$c_1 = c_0(2I - ac_0)$$

$$c_2 = c_1(2I - ac_1)$$

$$c_3 = c_2(2I - ac_2)$$

$$\cdots\cdots\cdots\cdots$$

$$c_m = c_{m-1}(2I - ac_{m-1}).$$

He obtains an upper bound for the difference between each element of c_m and the corresponding element of a^{-1} and indicates that the method is especially valuable when a good initial approximation is available and when a large number of decimal places is required. He suggests that a good initial approximation might be obtained by a direct method carried to a small number of places and later [L] that tables of the inverses of certain matrices might be used.

Hotelling [M] has also aided in the problem of solving the characteristic equation and in determining the modal columns or characteristic vectors by iterative methods. He has also given references to important work of other authors [N].

The reader is referred in addition to the articles by Hotelling, to the work of Aitken [O], and to the book by Frazer, Duncan, and Collar [P] for an extensive treatment of the basic iterative methods. The reader is also referred to a recent paper by Lanczos [Q.1] for a treatment of the general problem by improved methods and to recent papers by Jahn [Q.2] and Collar [Q.3] for the use of iterative methods as applied to problems involving the characteristic equation.

REFERENCES

A. 1. L. Guttman, "Enlargement methods for computing the inverse matrix," *Annals of Mathematical Statistics*, **17**, 336–343 (1946).
2. R. A. Frazer, W. J. Duncan, and A. R. Collar, *Elementary Matrices*, Cambridge University Press, Cambridge, 1947. See section 4.9.
3. J. Morris, *The Escalator Method in Engineering Vibration Problems*, Chapman and Hall, Ltd., London, 1947.

B. See [A.1], section 8.

C. See [A.2], pp. 115–116.

D. 1. M. Bocher, *Introduction to Higher Algebra*, The Macmillan Co., New York, 1907. See p. 33.
2. A. C. Aitken, *Determinants and Matrices*, Oliver and Boyd, Edinburgh, 1942. See section 42.

E. A. C. Aitken, "The evaluation of determinants, etc.," *Proceedings Edinburgh Mathematical Society*, Series 2, **3**, 207–219 (1932). See section 4.

F. 1. W. G. Cochran, "The omission or addition of an independent variate in multiple linear regression," *Supplement Journal Royal Statistical Society*, **5**, 171–176 (1938).
2. P. Boschan, "The consolidated Doolittle technique" (abstract), *Annals of Mathematical Statistics*, **17**, 503 (1946).

G. H. Hotelling, "Some new methods in matrix calculation," *Annals of Mathematical Statistics*, **14**, 1–34 (1943).

H. See E. Whittaker and G. Robinson, *The Calculus of Observations*, Blackie and Son, London, 1924, section 130.

I. See [G], section 5.

J. H. Cross, "Analysis of continuous frames by distributing fixed end moments," *Proceedings of the American Society of Civil Engineers* (1930).

K. See [A.3], Chapter 6.

L. H. Hotelling, "Practical problems of matrix calculation," *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, 1949, pp. 275–293. See section 5.

M. See [G], section 12.

N. See [G], pp. 33–34.

O. A. C. Aitken, "Studies in practical mathematics. II. The evaluation of the latent roots and latent vectors of a matrix," *Proceedings Royal Society, Edinburgh*, **57**, 269–304 (1937).

P. See [A.2], Chapter IV.

Q. 1. C. Lanczos, "An iteration method for the solution of eigenvalue problem of linear differential and integral operators." Unpublished manuscript.
2. H. A. Jahn, "Improvement of an approximate set of latent roots and modal columns of a matrix by methods akin to those of classical perturbation theory," *Quarterly Journal of Mechanics and Applied Mathematics*, **1**, 131–144 (1948).
3. A. R. Collar, "Some notes on Jahn's method for the improvement of approximate latent roots and vectors of a square matrix," *Quarterly Journal of Mechanics and Applied Mathematics*, **1**, 145–148 (1948).

EXERCISES

1. Calculate the inverse of the matrix of the coefficients of exercise 6.11, using the method of Table 16.2*a*.

2. Calculate the adjoint matrix of the coefficients of the problem of exercise 4.10, using the method of Table 16.2*b*.

3. Calculate the adjoint of

$$\begin{bmatrix} 1.0 & 0.2 & 0.3 & 0.4 \\ 0.2 & 1.0 & 0.4 & 0.6 \\ 0.3 & 0.4 & 1.0 & 0.7 \\ 0.4 & 0.6 & 0.7 & 1.0 \end{bmatrix}$$

using the method of Table 16.2*c*.

4. Calculate the adjugate of the matrix of coefficients of the problem of exercise 4.10, using T. Smith's method.

5. Calculate the inverse of the problem of exercise 3, using the method of Table 16.2*h*.

6. Calculate the inverse of the problem of Table 16.2*i* using the iterative method suggested by Hotelling.

7. Calculate the inverse of the problem of exercise 3, using the iterative method suggested by Hotelling.

I. See [G], section 5.

J. H. Cross, "Analysis of continuous frames by distributing fixed end moments," *Proceedings of the American Society of Civil Engineers* (1930).

K. See [A.3], Chapter 6.

L. H. Hotelling, "Practical problems of matrix calculation," *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, 1949, pp. 275–293. See section 5.

M. See [G], section 12.

N. See [G], pp. 33–34.

O. A. C. Aitken, "Studies in practical mathematics. II. The evaluation of the latent roots and latent vectors of a matrix," *Proceedings Royal Society, Edinburgh*, **57**, 269–304 (1937).

P. See [A.2], Chapter IV.

Q. 1. C. Lanczos, "An iteration method for the solution of eigenvalue problem of linear differential and integral operators." Unpublished manuscript.
2. H. A. Jahn, "Improvement of an approximate set of latent roots and modal columns of a matrix by methods akin to those of classical perturbation theory," *Quarterly Journal of Mechanics and Applied Mathematics*, **1**, 131–144 (1948).
3. A. R. Collar, "Some notes on Jahn's method for the improvement of approximate latent roots and vectors of a square matrix," *Quarterly Journal of Mechanics and Applied Mathematics*, **1**, 145–148 (1948).

EXERCISES

1. Calculate the inverse of the matrix of the coefficients of exercise 6.11, using the method of Table 16.2*a*.

2. Calculate the adjoint matrix of the coefficients of the problem of exercise 4.10, using the method of Table 16.2*b*.

3. Calculate the adjoint of

$$\begin{bmatrix} 1.0 & 0.2 & 0.3 & 0.4 \\ 0.2 & 1.0 & 0.4 & 0.6 \\ 0.3 & 0.4 & 1.0 & 0.7 \\ 0.4 & 0.6 & 0.7 & 1.0 \end{bmatrix}$$

using the method of Table 16.2*e*.

4. Calculate the adjugate of the matrix of coefficients of the problem of exercise 4.10, using T. Smith's method.

5. Calculate the inverse of the problem of exercise 3, using the method of Table 16.2*h*.

6. Calculate the inverse of the problem of Table 16.2*i* using the iterative method suggested by Hotelling.

7. Calculate the inverse of the problem of exercise 3, using the iterative method suggested by Hotelling.

CHAPTER 17

The Errors of Linear Computations

17.1 Introduction. We are now in a position to examine the question of the errors in the solutions of simultaneous equations when the coefficients are subject to error. A single solution may be used to represent the infinity of problems, as indicated in section 3.2, but then bounds for the error should be available. The methods of Chapter 2 can be used in computing bounds if range numbers or approximation-error numbers are used throughout the calculation. These numbers demand extensive computation, the resulting bounds may not be very close, and they may vary with the order of elimination. A more satisfactory method, as is pointed out in section 2.13, is the use of incomplete numbers if a formula is available for computing the bound for the incomplete number so that the result can be stated as a complete approximate number in approximation-error or in range form. The first-order approximation for the error of a determinant is presented in Chapter 10. Corresponding formulas for the error in a determinantal ratio, the errors in the solutions of simultaneous linear equations, the errors in the terms of the inverse matrix, etc., are presented in this chapter. This chapter also deals with the allied problems of the elimination of mistakes, the control of errors resulting from rounding off, the solution of adjusted equations, and similar problems.

17.2 Kinds of errors. Von Neumann and Goldstine [A] give a general discussion of the various types of error involved when a scientific problem is solved by numerical computation. The reader is referred to this article for a more extensive discussion. Since our objective is the solution of problems after they have been put in linear form, we need only consider two basic types of error:

(*a*) Those that result from the fact that the coefficients themselves may be subject to error.

(*b*) Those that result from the fact that approximate operations may be used in the process of solution.

The errors of type *a*, those that result from errors of measurement and the necessity of using digital numbers in setting up the problem,

are discussed in Chapter 1. Milne [B] calls these *inherent errors*. If the errors are not known exactly, it is customary to set some bounds for them. Such bounds can usually be determined from the value of the physical processes involved. If no such bounds can be determined, sets of hypothetical bounds can be taken so that the bounds for the errors of the result may be stated in terms of these different hypothetical bounds for the errors of the coefficients.

The errors of type b result from rounding off during the computational process. Strictly speaking, these errors do not include the rounding-off errors of the original statement of the problem that are errors of type a. Frequently we carry out solutions with incomplete numbers that have the number of places indicated by the errors of type a, and in this case the errors of type a and those of type b may be confused. However, the incomplete numbers may be carried to more places than are warranted by the type a errors, thus making the initial type b errors smaller. With use of suitable theory, the errors of type b can be controlled and bounds can be determined for the errors of type a. Exact methods, if they are practical, may be used to eliminate all type b errors. The study of the control of the accumulation of type b errors with approximate methods is sometimes known as *error control* [C].

The four basic types of linear problems discussed in section 3.2 may be stated, in the language of type a and type b errors, as:

(a) Problems involving neither type a nor type b errors.
(b) Problems involving type b errors only.
(c) Problems involving type a errors only.
(d) Problems involving type a and type b errors.

We should also consider the control of mistakes. This topic is discussed in the following sections since the elimination of mistakes is very important in linear computation.

17.3 The use of exact methods. Exact methods are most useful in error control and in eliminating mistakes. The forward solution by the method of determinants, for example, features an exact division that serves as a check on mistakes and indicates that there have been no rounding-off errors. Sometimes these exact methods lead to incomplete numbers beyond machine capacity in the forward solution, and in this case the exact divisibility feature is not present (unless the exact values are calculated by methods for operations beyond machine capacity). In any case each number can be computed to machine capacity so that for most problems the type b errors are very small compared to the type a errors.

A similar statement holds for the back solution since it is conventional, even with exact methods, to carry the back solution to a fixed number of decimal places. We can carry through the back solution to machine capacity so that the type b errors are small. For example, the nine-decimal-place solutions of Table 4.8a are easily obtained from the forward solution. The answers may then be rounded off to four decimal places, with the knowledge that type b errors have not accumulated to an amount, for each x_i, that is greater in absolute value than 0.00005. Thus the four-decimal-place answer of Table 4.8a that is obtained with a four-decimal-place back solution is the correct four-decimal-place back solution.

17.4 Use of sum checks. Most elimination methods have this property: the elimination process can be applied to the sum of certain quantities to obtain the sum of the results of the operations applied to the given quantities. This check is exact if the operations are exact and is approximate if the operations involve approximations. The row sum check, used in various tables, is an illustration of this. A column sum check could be used similarly. In calculations involving many variables, we may insert a sum check periodically, for instance, after every ten columns, to assist in the discovery of mistakes and to watch the cumulation of errors of type b.

If the method is exact, as in Table 4.8a, the check sum in the forward solution is exact. In general, mistakes, as well as errors of type b, are eliminated. The use of the sum check may also be carried through the back solution, as in Table 4.8a. Now the back solution consists in U_9 operations in which the denominator is exact. The absence of serious mistakes and the smallness of the cumulations of the type b errors are indicated when the decimal approximation for the back solution for sums differs from the decimal approximation for the solution by exactly unity for each x_i.

The situation is slightly different, as regards the forward solution, if approximate methods are used. However, the continued calculations of the values of a row sum column and the corresponding values of the sum of each row, as in Table 6.4a, enable us to examine for mistakes, and to compare for errors, at each step and at the end of the elimination process itself. Thus the comparison of 2.1748 with 2.1748 in the third from the last line of Table 6.4a indicates that there are no serious mistakes and that the errors of type b seem to play a minor role.

17.5 The final verification. The real test of a solution is whether or not it satisfies the original equations. If the equations are exact and the solution is exact, there are no errors present, and we can find by substitution if mistakes are present. If the equations are exact, but the

solutions are not, we can find by substitution a new set of equations with right side coefficients a'_{ij}, as in Table 6.4a, to compare with original equations. Close agreement indicates that there are no serious mistakes and that the errors of type b are trivial. The results of Table 6.4a show, for example, that the values

$$(1) \quad x_1 = -0.9366, \qquad x_2 = 0.0602, \qquad x_3 = 0.8152, \qquad x_4 = 1.1748$$

constitute an exact solution of

$$(2) \quad \begin{aligned} x_1 + 0.4x_2 + 0.5x_3 + 0.6x_4 &= 0.19996 \\ 0.4x_1 + x_2 + 0.3x_3 + 0.4x_4 &= 0.40004 \\ 0.5x_1 + 0.3x_2 + x_3 + 0.2x_4 &= 0.59992 \\ 0.6x_1 + 0.4x_2 + 0.2x_3 + x_4 &= 0.79996 \end{aligned}$$

and as such they are satisfactory approximations to the solution of (2), with the terms on the right replaced by 0.2, 0.4, 0.6, 0.8, respectively.

The situation is not so satisfactory when the coefficients are themselves subject to error for, as is shown in section 3.2, we have a multiple infinity of sets of equations. We may well select one of these sets of equations, such as (3.2.2), as a basic set whose solution may serve as a typical solution of all the sets. Or we may select a set that gives an extreme value for some x_i. In either case the final verification should be applied to a certain set of exact equations. Occasionally we may wish to know how well a specific solution satisfies any of the multiple infinity of solutions. For example, how well does the solution of

$$(3) \quad ax = f$$

satisfy

$$(4) \quad [a + \epsilon(a)]x = f + \epsilon(f)?$$

Here (4) is the matrix generalization of (3.2.1), with a and f the matrices of the approximation parts of the approximation-error coefficients and $\epsilon(a)$ and $\epsilon(f)$ the error parts. The values of $\epsilon(a)$ and $\epsilon(f)$ are known for any specific equations. Now if x is the solution of (3), the left side of (4) becomes

$$(5) \quad [a + \epsilon(a)]x = f + \epsilon(a)x.$$

The right side of (5) can then be compared with the right side of (4) or, what amounts to the same thing, $\epsilon(a)x$ can be compared with $\epsilon(f)$. If the absolute values of $\epsilon(a)x$ are smaller than those of $\epsilon(f)$, the solution of (3) may also be considered as a suitable solution of (4).

As an example, consider the illustration of Table 5.6. The exact solution of $ax = f$ is given by $x_1 = 2$, $x_2 = 1$, $x_3 = 3$, $x_4 = -2$. If the values $\epsilon(a)$ are given by

$$|\epsilon(a_{ij})| = |\epsilon_{ij}| \leq 0.0005 \tag{6}$$

and each sign is taken equal to that of the value of the X it multiplies, we get for the right side 23.000(4), 57.000(4), 48.000(4), $-68.000(4)$. The errors are less in absolute value than those of $\epsilon(f)$ if

$$|\epsilon(f_i)| = |\epsilon_i| \leq 0.005. \tag{7}$$

We may say then that the solution of Table 5.6*a* may be used in a general way to describe the multiple infinity of solutions that arise from equations whose coefficients on the left are limited to errors of 0.0005 and whose coefficients on the right are limited to errors of 0.005.

In general the solutions of exact determinate equations are unique. The solution can be determined to many places and the results can then be rounded off to obtain the required approximation. Of course a solution using incomplete numbers may vary from the correct answer because of the cumulation of type *b* errors. But in general there is only one result to the solution of $ax = f$, if a and f are exact, while a is nonsingular and f not zero, and that is $x = a^{-1}f$.

A somewhat different situation exists in regard to approximate equations, such as (4). Different solutions are possible, depending on the values of $\epsilon(a)$ and $\epsilon(f)$, and any one of these solutions might properly be called a solution of (4). Indeed a solution of (4) might be defined as any set of values that satisfies (4) within the limits specified. Such solutions, of course, need not be unique. Claim could be made that the solution of a set of approximate equations has been found when one set of values, which satisfies the equations within the limits specified, is available.

17.6 The errors of determinants. The evaluation of determinants with approximate numbers as elements is discussed in section 10.12. A method using approximation-error numbers is followed by a more satisfactory method using incomplete numbers with the use of an auxiliary formula (10.12.11) for finding the maximum first-order error of the determinant. Illustration to determinants of the third order is presented in Chapter 10.

We are now in a position to obtain the maximum values of the errors of determinants of higher order since the values of the inverse and the adjoint are available by the methods of Chapters 13 and 14. The

formulas of Chapter 10 may be written as

$$\epsilon(\Delta) = \sum\sum \epsilon_{ij} A_{ij}$$

(1)
$$\eta(\Delta) = \sum_i \sum_j \eta_{ij} \mid A_{ij} \mid$$

$$\eta(\Delta) = \eta \sum_i \sum_j \mid A_{ij} \mid$$

when the terms A_{ij} are the terms of the adjoint of the matrix having the determinant Δ. Similarly,

$$\epsilon(\Delta) = \Delta \sum_i \sum_j \epsilon_{ij} c_{ij}$$

(2)
$$\eta(\Delta) = \mid \Delta \mid \sum_i \sum_j \eta_{ij} \mid c_{ij} \mid$$

$$\eta(\Delta) = \mid \Delta \mid \eta \sum_i \sum_j \mid c_{ij} \mid .$$

To illustrate (1), consider the determinant of the matrix of Table 14.5*e*. If $\mid \epsilon_{ij} \mid \leq 0.00005$, it follows that

$$\eta(\Delta) = (0.00005) \sum_i \sum_j \mid A_{ij} \mid$$

so that

$$\eta(\Delta) = (0.00005)(599678)$$

$$\eta(\Delta) = 29.9839$$

and the value of the determinant may be written

$$\Delta = 2{,}305{,}327(30).$$

If the elements of the matrix are significant numbers, $\eta = 0.5$, and we have

$$\Delta = 2{,}305{,}327(299839)$$

and the relative error may be as much as 13%.

We use the determinant of the matrix of Table 13.6*a* to illustrate (2). The approximate value of Δ is 0.3660. The approximate value of $\sum_i \sum_j \mid c_{ij} \mid$ is 12.0973. If $\mid \epsilon_{ij} \mid \leq 0.00005$, we have

$$\eta(\Delta) = (0.3660)(0.00005)(12.0973)$$

$$= 0.000221$$

so that

$$\Delta = 0.3660(2).$$

The formulas used in this section contain only first-order terms. They are especially applicable to the usual case in which the maximum relative errors are small.

A more extensive treatment of the errors of determinants, including second-order errors and statistical or probability formulas, is given by Etherington [D].

17.7 The solution of linear equations with coefficients subject to error. We have the basic matrix equation

$$[a + \epsilon(a)][x + \epsilon(x)] = f + \epsilon(f) \tag{1}$$

when the coefficients of the linear equations are subject to error. Here a and f are the matrices of the approximation part of the approximation-error members and $\epsilon(a)$ and $\epsilon(f)$ are the corresponding errors. We wish to find the values of $\epsilon(x)$ when a and $a + \epsilon(a)$ are non-singular and x is the solution of the exact matrix equation.

$$ax = f. \tag{2}$$

Now $x = a^{-1}f$ can either be determined exactly or the approximate solution can be carried to a sufficient number of places so that its difference from the true $a^{-1}f$ is negligible when compared with $a^{-1}\epsilon(a)$ and $a^{-1}\epsilon(f)$. We then expand (1) and get

$$ax + a\epsilon(x) + \epsilon(a)x + \epsilon(a)\epsilon(x) = f + \epsilon(f). \tag{3}$$

Now we subtract (2) from (3) to get

$$a\epsilon(x) + \epsilon(a)x + \epsilon(a)\epsilon(x) = \epsilon(f). \tag{4}$$

We further limit our approach to the use of first-order errors. In this case $\epsilon(a)\epsilon(x)$ may be neglected and we have

$$a\epsilon(x) + \epsilon(a)x = \epsilon(f). \tag{5}$$

We transfer $\epsilon(a)x$ to the right side to get

$$a\epsilon(x) = \epsilon(f) - \epsilon(a)x \tag{6}$$

from whence

$$\begin{aligned} \epsilon(x) &= a^{-1}[\epsilon(f) - \epsilon(a)x] = c[\epsilon(f) - \epsilon(a)x] \\ &= c[\epsilon(f) - \epsilon(a)a^{-1}f] = c[\epsilon(f) - \epsilon(a)cf]. \end{aligned} \tag{7}$$

This formula, which gives the first-order approximation to x when a, $\epsilon(a)$, f, and $\epsilon(f)$ are specified, is the basic formula of this section.

We may substitute in (7) at once, after obtaining the value of $c = a^{-1}$, if specific values of a, $\epsilon(a)$, f, and $\epsilon(f)$ are given. In typical problems,

however, the values of $\epsilon(a)$ and $\epsilon(f)$ are not known, but a bound for each is known. In this case we have

$$\eta(x) = |\,c\,|[\eta(f) + \eta(a)|\,x\,|]. \tag{8}$$

If, in addition, each element of $\eta(f)$ and $\eta(a)$ is less than η, we can write

$$\eta(x) = \eta|\,c\,|[(1_f) + (1_a)|\,x\,|] \tag{9}$$

where (1_f) and (1_a) are matrices of unit terms with the number of rows and columns of f and a, respectively. We denote $|\,c\,|$ by C and $|\,x\,|$ by X and have

$$\eta(x) = \eta C[(1_f) + (1_a)X]. \tag{10}$$

The detail of (10) is written, when $p = 3$, as

$$\eta(x) = \eta \begin{bmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{31} & C_{32} & C_{33} \end{bmatrix} \left\{ \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \right\} \tag{11}$$

so that

$$\begin{aligned} \eta(x_1) &= \eta(C_{11} + C_{12} + C_{13})(1 + \Sigma X_i) \\ \eta(x_2) &= \eta(C_{21} + C_{22} + C_{23})(1 + \Sigma X_i) \\ \eta(x_3) &= \eta(C_{31} + C_{32} + C_{33})(1 + \Sigma X_i). \end{aligned} \tag{12}$$

Similarly, the general formula (10) becomes

$$\eta(x_k) = \eta \sum_j C_{kj}(1 + \Sigma X_i) = \eta(1 + \Sigma X_i) \sum_j C_{kj} \tag{13}$$

and it follows that the values of maximum errors of x_i are proportional to the sum of the absolute values of the elements in the kth row of the inverse of a.

Since the elements of C are the elements of the absolute values of the adjoint divided by Δ, we have

$$C_{kj} = \frac{|\,A_{kj}\,|}{\Delta} \tag{14}$$

and (13) becomes

$$\eta(X_k) = \frac{\eta(1 + \Sigma X_i) \sum_j A_{kj}}{\Delta}. \tag{15}$$

The formulas (13) and (15) are immediately applicable to problems in which the solutions and the inverse or adjoint are known. Because the solutions of the earlier chapters feature the transpose of the inverse and adjoint, we need the sums of columns rather than rows. Where the

solution is presented as a row vector and the transpose of the inverse or adjoint is presented directly below it, the values of the columns correspond to the values of x_i above. This is illustrated in Table 17.7*a*,

TABLE 17.7*a*

THE SOLUTIONS AND THE TRANSPOSE OF THE ADJOINT

General				Illustration			
x_1	x_2	x_3	x_4	2	1	3	−2
A'_{11}	A'_{12}	A'_{13}	A'_{14}	66233	−16033	42069	−6503
A'_{21}	A'_{22}	A'_{23}	A'_{24}	56151	28558	33194	−52258
A'_{31}	A'_{32}	A'_{33}	A'_{34}	−53068	36236	18224	45899
A'_{41}	A'_{42}	A'_{43}	A'_{44}	−35013	9659	−47056	53524
$\Sigma\lvert A'_{i1}\rvert$	$\Sigma\lvert A'_{i2}\rvert$	$\Sigma\lvert A'_{i3}\rvert$	$\Sigma\lvert A'_{i4}\rvert$	210465	90486	140543	158184

where the adjugate is used. A general presentation is given on the left, and a particular illustration on the right. The solution is that of the problem of Table 5.6 while the adjugate of the coefficients is obtained from Table 14.5*e*. The sum of the absolute values of the elements of the columns is also exhibited.

Now $1 + \Sigma X_i = 1 + 2 + 1 + 3 + 2 = 9$. We need only the values of Δ and η to substitute in (15). $\Delta = 2{,}305{,}327$ from Table 5.6*a* and Table 14.5*e*. If $\eta = 0.005$ (that is, all the coefficients are correct to two decimal places), we have

$$\frac{\eta(1 + \Sigma X)}{\Delta} = \frac{0.005(9)}{2{,}305{,}327} = 1.952.10^{-8}.$$

It follows that

$$\eta(x_1) = 210465(1952.10^{-8}) = 0.0041$$

$$\eta(x_2) = 90486(1.952.10^{-8}) = 0.0018$$

$$\eta(x_3) = 140543(1.952.10^{-8}) = 0.0027$$

$$\eta(x_4) = 158184(1.952.10^{-8}) = 0.0031$$

and that the values of the solutions of the equations of Table 5.6*a*, with coefficients correct to two decimal places, may be indicated by

$$x_1 = 2.000(4), \quad x_2 = 1.000(2), \quad x_3 = 3.000(3), \quad x_4 = -2.000(3).$$

The solution of a problem using the transpose of the inverse matrix is similarly presented in Table 17.7*b*. The values of x_i for the problem

TABLE 17.7*b*

ERRORS OF x_i WITH USE OF INVERSE MATRIX

−0.9366	0.0602	0.8152	1.1748	3.0868
2.0708	−0.1913	−0.7759	−1.0107	0.005
−0.1913	1.2842	−0.2185	−0.3552	0.019934
−0.7759	−0.2185	1.3988	0.2731	
−1.0107	−0.3552	0.2731	1.6941	
4.0487	2.0492	2.6663	3.3331	
0.0807	0.0408	0.0532	0.0664	

are obtained from Table 6.4*a*, and the approximate elements of the inverse are obtained from Table 13.6*a*. The value $1 + X_1 + X_2 + X_3 + X_4$, placed at the right of the first row, is multiplied by $\eta = 0.005$ to get 0.019934. This is multiplied in turn by the sums of absolute values to obtain the maximum errors.

The values of the first row of the table give the approximation part; the values of the last row of the table give the error part if the coefficients are correct to two decimal places. A comparison of these results leads us to state the approximate answer:

$$x_1 = -0.94(8), \qquad x_2 = 0.06(4), \qquad x_3 = 0.82(5), \qquad x_4 = 1.17(6).$$

The formulas (13) and (15) do give first-order approximations to the maximum errors of the x_i, but they do not constitute a solution since the maximum errors of x_i are not necessarily associated with the maximum errors of x_j. We may obtain, however, the values of the errors of x_i that are associated with a maximum error for any specific x_j. Simultaneously with this we may obtain the errors of the coefficients of the equation so as to produce equations having these maximum errors. To do this it is wise to return to (7) and to select each error of $\epsilon(f)$ and $\epsilon(a)$ so as to maximize the different values of $\epsilon(x)$. Thus for $p = 2$, we have

$$(16) \qquad \begin{bmatrix} \epsilon(x_1) \\ \epsilon(x_2) \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} \left\{ \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix} - \begin{bmatrix} \epsilon_{11} & \epsilon_{12} \\ \epsilon_{21} & \epsilon_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \right\}$$

so that

$$(17) \qquad \epsilon(x_1) = c_{11}\epsilon_1 - c_{11}\epsilon_{11}x_1 - c_{11}\epsilon_{12}x_2 + c_{12}\epsilon_2 - c_{12}\epsilon_{21}x_1 - c_{12}\epsilon_{22}x_2$$

The solution of a problem using the transpose of the inverse matrix is similarly presented in Table 17.7*b*. The values of x_i for the problem

TABLE 17.7*b*

ERRORS OF x_i WITH USE OF INVERSE MATRIX

−0.9366	0.0602	0.8152	1.1748	3.9868
2.0708	−0.1913	−0.7759	−1.0107	0.005
−0.1913	1.2842	−0.2185	−0.3552	0.019934
−0.7759	−0.2185	1.3988	0.2731	
−1.0107	−0.3552	0.2731	1.6941	
4.0487	2.0492	2.6663	3.3331	
0.0807	0.0408	0.0532	0.0664	

are obtained from Table 6.4*a*, and the approximate elements of the inverse are obtained from Table 13.6*a*. The value $1 + X_1 + X_2 + X_3 + X_4$, placed at the right of the first row, is multiplied by $\eta = 0.005$ to get 0.019934. This is multiplied in turn by the sums of absolute values to obtain the maximum errors.

The values of the first row of the table give the approximation part; the values of the last row of the table give the error part if the coefficients are correct to two decimal places. A comparison of these results leads us to state the approximate answer:

$$x_1 = -0.94(8), \quad x_2 = 0.06(4), \quad x_3 = 0.82(5), \quad x_4 = 1.17(6).$$

The formulas (13) and (15) do give first-order approximations to the maximum errors of the x_i, but they do not constitute a solution since the maximum errors of x_i are not necessarily associated with the maximum errors of x_j. We may obtain, however, the values of the errors of x_i that are associated with a maximum error for any specific x_j. Simultaneously with this we may obtain the errors of the coefficients of the equation so as to produce equations having these maximum errors. To do this it is wise to return to (7) and to select each error of $\epsilon(f)$ and $\epsilon(a)$ so as to maximize the different values of $\epsilon(x)$. Thus for $p = 2$, we have

$$(16) \qquad \begin{bmatrix} \epsilon(x_1) \\ \epsilon(x_2) \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} \left\{ \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix} - \begin{bmatrix} \epsilon_{11} & \epsilon_{12} \\ \epsilon_{21} & \epsilon_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \right\}$$

so that

$$(17) \quad \epsilon(x_1) = c_{11}\epsilon_1 - c_{11}\epsilon_{11}x_1 - c_{11}\epsilon_{12}x_2 + c_{12}\epsilon_2 - c_{12}\epsilon_{21}x_1 - c_{12}\epsilon_{22}x_2$$

and

$$(18)\quad \epsilon(x_2) = c_{21}\epsilon_1 - c_{21}\epsilon_{11}x_1 - c_{21}\epsilon_{12}x_2 + c_{22}\epsilon_2 - c_{22}\epsilon_{21}x_1 - c_{22}\epsilon_{22}x_2.$$

To maximize $\epsilon(x_1)$ take the signs of ϵ_1 and ϵ_2 identical with those of c_{11} and c_{12}, respectively; take the signs of ϵ_{11}, ϵ_{12}, ϵ_{21}, and ϵ_{22} identical with those of $-c_{11}x_1$, $-c_{11}x_2$, $-c_{12}x_1$, and $-c_{12}x_2$, respectively. These values when substituted in (18) result in a value of $\epsilon(x_2)$, which is not necessarily a maximum. To maximize $\epsilon(x_2)$ assign the values of the signs of ϵ_1, ϵ_2, ϵ_{11}, ϵ_{12}, ϵ_{21}, and ϵ_{22} to coincide with those of c_{21}, c_{22}, $-c_{21}x_1$, $-c_{21}x_2$, $-c_{22}x_1$, and $-c_{22}x_2$, respectively. The value of $\epsilon(x_1)$ may then be computed for this second set of values of the errors. The determination of the signs of ϵ_i and ϵ_{ij} and the calculation of the maximum error of x_i is facilitated by using tables similar to Tables 17.7*c* and 17.7*d*.

TABLE 17.7*c*

CALCULATION OF THE MAXIMUM $\epsilon(x_1)$

x_1	x_2	
$-c_{11}x_1$	$-c_{11}x_2$	c_{11}
$-c_{12}x_1$	$-c_{12}x_2$	c_{12}

The signs of the elements of the table are the proper signs of the corresponding errors. The maximum value of $\epsilon(x_1)$ may then be computed by multiplying each value in the table by the corresponding error with its appropriate sign and by summing the results. The bound for the error, rather than the error itself, may be used in the situation in which a bound for each error is known. In this case a bound for the error of the solution is obtained.

A similar treatment for the calculation of the maximum $\epsilon(x_2)$ is based on Table 17.7*d*.

TABLE 17.7*d*

CALCULATION OF THE MAXIMUM $\epsilon(x_2)$

x_1	x_2	
$-c_{21}x_1$	$-c_{21}x_2$	c_{21}
$-c_{22}x_1$	$-c_{22}x_2$	c_{22}

By using Table 17.7*c* and Table 17.7*d* we may calculate the value of $\epsilon(x_2)$ for the errors that maximize $\epsilon(x_1)$, or we may calculate the value of $\epsilon(x_1)$ for the errors that maximize $\epsilon(x_2)$. In the first case we substitute in (18), using the errors that maximize (17); in the second case we substitute in (17), using the errors that maximize (18).

We use an illustration in which the bound for each error is 0.01. Here the values of ϵ may be replaced by ± 1 with a final multiplication by $\eta = 0.01$. The problem is to solve the approximate equations

$$\begin{aligned} x_1 + x_2 &= 3 \\ 2x_1 + 3x_2 &= 7, \end{aligned} \tag{19}$$

where each of the coefficients may have an error as large as 0.01 in absolute value.

The solution is presented in Table 17.7*e*. The compact method of single division is used to find the values of x_1 and x_2 and the elements of the transpose of the inverse. Values corresponding to those of Tables 17.7*c* and 17.7*d* are there exhibited. A general algebraic presentation is followed by the application to (19).

The compact method of single division, for this particular illustration, results in exact values of x_i and c_{ij}, so that there are no type b errors and the type a errors are our only concern.

The maximum value of $\epsilon(x_1)$ is easily computed since it is

$$\begin{aligned} \eta(x_1) &= 0.01[(-6)(-1) + (-3)(-1) + (3)(1) + (2)(1) \\ &\qquad + (1)(1) + (-1)(-1)] \\ &= 0.16 \end{aligned} \tag{20}$$

whereas the corresponding value of $\eta(x_2)$ is

$$\begin{aligned} \eta(x_2) &= 0.01[(4)(-1) + (2)(-1) + (-2)(1) + (-2)(1) \\ &\qquad + (-1)(1) + (1)(-1)] \\ &= -0.12. \end{aligned} \tag{21}$$

Similarly, we have for the maximum value of $\epsilon(x_2)$

$$\begin{aligned} \eta(x_2) &= 0.01[(4)(1) + (2)(1) + (-2)(-1) + (-2)(-1) \\ &\qquad + (-1)(-1) + (1)(1)] \\ &= 0.12 \end{aligned} \tag{22}$$

$$\begin{aligned} \eta(x_1) &= 0.01[(-6)(1) + (-3)(1) + (3)(-1) + (2)(-1) \\ &\qquad + (1)(-1) + (-1)(1)] \\ &= -0.16. \end{aligned} \tag{23}$$

TABLE 17.7e

SOLUTION OF SIMULTANEOUS EQUATIONS WITH COEFFICIENTS SUBJECT TO ERROR

x_1	x_2				
a_{11} a_{21}	a_{12} a_{22}	a_{13} a_{23}	1 0	0 1	
a_{11} a_{21}	b_{12} $g_{22\cdot1}$	b_{13} $b_{23\cdot1}$	$1/a_{11}$ $-a_{21}/a_{11}g_{22\cdot1}$	0 $1/g_{22\cdot1}$	
x_1	x_2				
c_{11}	c_{21}				
c_{12}	c_{22}				
$-c_{11}x_1$ $-c_{12}x_1$	$-c_{11}x_2$ $-c_{12}x_2$	c_{11} c_{12}	$\epsilon(x_1)$ $\epsilon(x_2)$	$x_1 + \epsilon(x_1)$ $x_2 + \epsilon(x_2)$	for max $\epsilon(x_1)$
$-c_{21}x_1$ $-c_{22}x_1$	$-c_{21}x_2$ $-c_{22}x_2$	c_{21} c_{22}	$\epsilon(x_1)$ $\epsilon(x_2)$	$x_1 + \epsilon(x_1)$ $x_2 + \epsilon(x_2)$	for max $\epsilon(x_2)$

ILLUSTRATION

x_1	x_2				
1 2	1 3	3 7	1 0	0 1	
1 2	1 1	3 1	1 −2	0 1	
2	1				
3	−2				
−1	1				
−6 2	−3 1	3 −1	0.16 −0.12	2.16 0.88	for max $\epsilon(x_1)$
4 −2	2 −1	−2 1	−0.16 0.12	1.84 1.12	for max $\epsilon(x_2)$

We are also in a position to produce the equations that have the extreme solutions

$$x_1 = 2.16, \quad x_2 = 0.88 \quad \text{and} \quad x_1 = 1.84, \quad x_2 = 1.12.$$

Adding the individual errors used in (20) and (21) to (19), we get

$$\begin{aligned} 0.99x_1 + 0.99x_2 &= 3.01 \\ 2.01x_1 + 3.01x_2 &= 6.99 \end{aligned} \tag{24}$$

which have the two-decimal-place solution $x_1 = 2.16$, $x_2 = 0.88$. Similarly we add the individual errors of (22) and (23) to (19) to form the equations

$$\begin{aligned} 1.01x_1 + 1.01x_2 &= 2.99 \\ 1.99x_1 + 2.99x_2 &= 7.01 \end{aligned} \tag{25}$$

which have the two-decimal-place solution $x_1 = 1.84$, $x_2 = 1.12$.

The theory is applied to a three-variable problem in Tables 17.7*f* and 17.7*g*. The method of determinants is used in getting the solution and the inverse. An algebraic presentation is given in Table 17.7*f*, and the scheme is applied to a numeric problem in Table 17.7*g*. The illustration is the one that Burington [E] has used to present the conventional determinantal methods of solving equations. The adjugate, rather than the transpose of the inverse, is featured. The bound for the error of each coefficient is 0.01. The method follows the general plan of Table 17.7*e*. The values of $k\eta(x_1)$ and $k\eta(x_2)$ when $\eta(x_1)$ is a maximum are

$$\begin{aligned} k\eta(x_1) = {} & (-5)(-1) + (10)(1) + (-15)(-1) + (5)(1) \\ & + (-1)(-1) + (2)(1) + (-3)(-1) + (1)(1) \\ & + (-3)(-1) + (6)(1) + (-9)(-1) + (3)(1) \\ = {} & 63 \end{aligned} \tag{26}$$

and

$$\begin{aligned} k\eta(x_2) = {} & (4)(-1) + (-8)(1) + (12)(-1) + (-4)(1) \\ & + (-4)(-1) + (8)(1) + (-12)(-1) + (4)(1) \\ & + (4)(-1) + (-8)(1) + (12)(-1) + (-4)(1) \\ = {} & -28. \end{aligned} \tag{27}$$

The values of $\eta(x_i)$ are obtained from the values of $k\eta(x_i)$ by multiplication by η and division by Δ. In this case $\eta = 0.01$ and $\Delta = 8$.

TABLE 17.7*f*

MAXIMUM ERRORS WITH THE METHOD OF DETERMINANTS

x_1	x_2	x_3				
a_{11}	a_{12}	a_{13}	a_{14}	Δ	0	0
a_{21}	a_{22}	a_{23}	a_{24}	0	Δ	0
a_{31}	a_{32}	a_{33}	a_{34}	0	0	Δ
	$d_{22\cdot1}$	$d_{23\cdot1}$	$d_{24\cdot1}$	*	*	0
	$d_{32\cdot1}$	$d_{33\cdot1}$	$d_{34\cdot1}$	*	*	*
		$d_{33\cdot12}$	$d_{34\cdot12}$	*	*	*

x_1	x_2	x_3				Max x_i	Min x_i	x_i
A'_{11}	A'_{12}	A'_{13}						
A'_{21}	A'_{22}	A'_{23}						
A'_{31}	A'_{32}	A'_{33}						
$-A'_{11}x_1$	$-A'_{11}x_2$	$-A'_{11}x_3$	A'_{11}	$k\eta(x_1)$	$\eta(x_1)$	x_1	x_1	
$-A'_{21}x_1$	$-A'_{21}x_2$	$-A'_{21}x_3$	A'_{21}	$k\eta(x_2)$	$\eta(x_2)$	x_2	x_2	x_1
$-A'_{31}x_1$	$-A'_{31}x_2$	$-A'_{31}x_3$	A'_{31}	$k\eta(x_3)$	$\eta(x_3)$	x_3	x_3	
$-A'_{12}x_1$	$-A'_{12}x_2$	$-A'_{12}x_3$	A'_{12}	$k\eta(x_1)$	$\eta(x_1)$	x_1	x_1	
$-A'_{22}x_1$	$-A'_{22}x_2$	$-A'_{22}x_3$	A'_{22}	$k\eta(x_2)$	$\eta(x_2)$	x_2	x_2	x_2
$-A'_{32}x_1$	$-A'_{32}x_2$	$-A'_{32}x_3$	A'_{32}	$k\eta(x_3)$	$\eta(x_3)$	x_3	x_3	
$-A'_{13}x_1$	$-A'_{13}x_2$	$-A'_{13}x_3$	A'_{13}	$k\eta(x_1)$	$\eta(x_1)$	x_1	x_1	
$-A'_{23}x_1$	$-A'_{23}x_2$	$-A'_{23}x_3$	A'_{23}	$k\eta(x_2)$	$\eta(x_2)$	x_2	x_2	x_3
$-A'_{33}x_1$	$-A'_{33}x_2$	$-A'_{33}x_3$	A'_{33}	$k\eta(x_3)$	$\eta(x_3)$	x_3	x_3	

As in the previous illustration, we can write the equations that have the extreme solutions since we know the proper signs for all the errors. The three sets of equations that give the maximum values of x_1, x_2, and x_3, respectively, are

$$\begin{aligned} 1.99x_1 + 1.01x_2 - 2.01x_3 &= -5.99 \\ 0.99x_1 + 1.01x_2 + 0.99x_3 &= 2.01 \\ -1.01x_1 - 1.99x_2 + 2.99x_3 &= 12.01 \end{aligned} \tag{28}$$

$$2.01x_1 + 0.99x_2 - 1.99x_3 = -6.01$$

(29) $$0.99x_1 + 1.01x_2 + 0.99x_3 = 2.01$$

$$-0.99x_1 - 2.01x_2 + 3.01x_3 = 11.99$$

and

$$2.01x_1 + 0.99x_2 - 1.99x_3 = -6.01$$

(30) $$0.99x_1 + 1.01x_2 + 0.99x_3 = 2.01$$

$$-1.01x_1 - 1.99x_2 + 2.99x_3 = 12.01.$$

The solutions of (28), (29), and (30) were obtained with the method of

TABLE 17.7*g*

MAXIMUM ERRORS WITH METHOD OF DETERMINANTS—ILLUSTRATION

x_1	x_2	x_3						
2	1	−2	−6	8	0	0		
1	1	1	2	0	8	0		
−1	−2	3	12	0	0	8		
	1	4	10	−8	16	0		
	−3	4	18	8	0	16		
		8	24	−8	24	8		
1	−2	3						
5	−4	−1						
1	4	3						
3	−4	1				Max x_i	Min x_i	x_i
−5	10	−15	5	63	0.079	1.079	0.921	
−1	2	−3	1	−28	−0.035	−2.035	−1.965	x_1
−3	6	−9	3	21	0.026	3.026	2.974	
4	−8	12	−4	−49	−0.061	0.939	1.061	
−4	8	−12	4	84	0.105	−1.895	−2.105	x_2
4	−8	12	−4	21	0.026	3.026	2.974	
1	−2	3	−1	−7	−0.009	0.991	1.009	
−3	6	−9	3	28	0.035	−1.965	−2.035	x_3
−1	2	−3	1	35	0.044	3.044	2.956	

determinants. The actual three-decimal-place answers are compared with the computed results of Table 17.7*g* in Table 17.7*h*.

TABLE 17.7*h*

COMPARISON OF SOLUTIONS

		x_1	x_2	x_3
maximum x_1	actual solution	1.080	−2.036	3.027
	computed solution	1.079	−2.035	3.026
maximum x_2	actual solution	0.940	−1.897	3.026
	computed solution	0.939	−1.895	3.026
maximum x_3	actual solution	0.991	−1.965	3.044
	computed solution	0.992	−1.965	3.044

Three other sets of extreme solutions can be obtained by using the errors that result in the minimum values of x_i. In this case each ϵ_i and ϵ_{ij} is given the sign opposite to that used above. These values of ϵ_{ij} are added to the values of a_{ij} to give the symbolic equations. Their solutions are found by adding the negatives of the values of $\epsilon(x_2)$ of Table 17.7*g* to the values of x_i.

The type of solution indicated is flexible enough to make possible the calculation of the maximum errors for other bounds. Thus if the bound for each error is 0.005, that is, the coefficients are significant to three places (or to two decimal places), the errors of Table 17.7*g* become 0.040, −0.018, 0.013, −0.031, 0.053, 0.013, −0.005, 0.018, and 0.022.

The compact method of single division may be used as the basis of the calculation of the solution and the inverse matrix. In this case we should carry the incomplete numbers to more places than are indicated by the type *a* errors so as not to confuse the type *b* and type *a* errors. An algebraic presentation of the general case for a three-variable problem is given in Table 17.7*i*.

The work of computing the $\epsilon(x_i)$ for a maximum or minimum x_i becomes tedious if the problem involves many variables. However, a technique can be worked out that is shorter if we are willing to dispense with the feature that exhibits the actual equations having the extreme solutions indicated. When $p = 3$, we know that

$$\epsilon(x_1) = c_{11}(\epsilon_1 - \epsilon_{11}x_1 - \epsilon_{12}x_2 - \epsilon_{13}x_3) + c_{12}(\epsilon_2 - \epsilon_{21}x_1 - \epsilon_{22}x_2 - \epsilon_{23}x_3) + c_{13}(\epsilon_3 - \epsilon_{31}x_1 - \epsilon_{32}x_2 - \epsilon_{33}x_3) \tag{31}$$

TABLE 17.7*i*

MAXIMUM ERRORS WITH THE COMPACT METHOD OF SINGLE DIVISION

x_1	x_2	x_3					
a_{11}	a_{12}	a_{13}	a_{14}	1	0	0	
a_{21}	a_{22}	a_{23}	a_{24}	0	1	0	
a_{31}	a_{32}	a_{33}	a_{34}	0	0	1	
a_{11}	b_{12}	b_{13}	b_{14}	*	0	0	
a_{21}	$g_{22\cdot1}$	$b_{23\cdot1}$	$b_{24\cdot1}$	*	*	0	
a_{31}	$g_{32\cdot1}$	$g_{33\cdot12}$	$b_{34\cdot12}$	*	*	*	
x_1	x_2	x_3					
c_{11}	c_{21}	c_{31}					
c_{12}	c_{22}	c_{32}					
c_{13}	c_{23}	c_{33}			Max x_i	Min x_i	x_i
$-c_{11}x_1$	$-c_{11}x_2$	$-c_{11}x_3$	c_{11}	$\eta(x_1)$	x_1	x_1	
$-c_{12}x_1$	$-c_{12}x_2$	$-c_{12}x_3$	c_{12}	$\eta(x_2)$	x_2	x_2	x_1
$-c_{13}x_1$	$-c_{13}x_2$	$-c_{13}x_3$	c_{13}	$\eta(x_3)$	x_3	x_3	
$-c_{21}x_1$	$-c_{21}x_1$	$-c_{21}x_3$	c_{21}	$\eta(x_1)$	x_1	x_1	
$-c_{22}x_1$	$-c_{22}x_2$	$-c_{22}x_3$	c_{22}	$\eta(x_2)$	x_2	x_2	x_2
$-c_{23}x_1$	$-c_{23}x_2$	$-c_{23}x_3$	c_{23}	$\eta(x_3)$	x_3	x_3	
$-c_{31}x_1$	$-c_{31}x_2$	$-c_{31}x_3$	c_{31}	$\eta(x_1)$	x_1	x_1	
$-c_{32}x_1$	$-c_{32}x_2$	$-c_{32}x_3$	c_{32}	$\eta(x_2)$	x_2	x_2	x_3
$-c_{33}x_1$	$-c_{33}x_2$	$-c_{33}x_3$	c_{33}	$\eta(x_3)$	x_3	x_3	

whereas the corresponding value of $\epsilon(x_2)$ is

$$\epsilon(x_2) = c_{21}(\epsilon_1 - \epsilon_{11}x_1 - \epsilon_{12}x_2 - \epsilon_{13}x_3) + c_{22}(\epsilon_2 - \epsilon_{21}x_1 - \epsilon_{22}x_2 - \epsilon_{23}x_3) + c_{23}(\epsilon_3 - \epsilon_{31}x_1 - \epsilon_{32}x_2 - \epsilon_{33}x_3), \tag{32}$$

where the ϵ_i and ϵ_{ij} of (32) are not in general those of (31). Now (31) can be written more symbolically as

$$\epsilon(x_1) = c_{11}E_{11} + c_{12}E_{21} + c_{13}E_{31}, \tag{33}$$

where

$$E_{i1} = \epsilon_i - \epsilon_{i1}x_1 - \epsilon_{i2}x_2 - \epsilon_{i3}x_3, \qquad \text{for} \qquad \epsilon(x_1). \tag{34}$$

Similarly, (32) may be written as

$$\epsilon(x_2) = c_{21}E_{12} + c_{22}E_{22} + c_{23}E_{32}, \tag{35}$$

where

$$E_{i2} = \epsilon_i - \epsilon_{i1}x_1 - \epsilon_{i2}x_2 - \epsilon_{i3}x_3, \qquad \text{for} \qquad \epsilon(x_2). \tag{36}$$

We now take the values of the ϵ_i and ϵ_{ij} in (31) and (33) so as to maximize $\epsilon(x_1)$. Then the use of these same values in (32) and (33) will give the value of $\epsilon(x_2)$ for max $\epsilon(x_1)$, and we may write

$$\epsilon(x_2)_{\max x_1} = \epsilon(x_2)_{x_1} = c_{21}E_{11} + c_{22}E_{21} + c_{23}E_{31}. \tag{37}$$

In a similar fashion, if the E_{i2} of (35) are determined so as to maximize $\epsilon(x_2)$, we have

$$\epsilon(x_2)_{\max x_1} = \epsilon(x_2)_{x_1} = c_{21}E_{11} + c_{22}E_{21} + c_{23}E_{31}. \tag{38}$$

The results (37) and (38) can be generalized into the matrix formula

$$\begin{bmatrix} \epsilon(x_1)_{x_1} & \epsilon(x_1)_{x_2} & \epsilon(x_1)_{x_3} \\ \epsilon(x_2)_{x_1} & \epsilon(x_2)_{x_2} & \epsilon(x_2)_{x_3} \\ \epsilon(x_3)_{x_1} & \epsilon(x_3)_{x_2} & \epsilon(x_3)_{x_3} \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix} \begin{bmatrix} E_{11} & E_{12} & E_{13} \\ E_{21} & E_{22} & E_{23} \\ E_{31} & E_{32} & E_{33} \end{bmatrix}. \tag{39}$$

Taking the transpose of each side of (39), we get

$$\begin{bmatrix} \epsilon(x_1)_{x_1} & \epsilon(x_2)_{x_1} & \epsilon(x_3)_{x_1} \\ \epsilon(x_1)_{x_2} & \epsilon(x_2)_{x_2} & \epsilon(x_3)_{x_2} \\ \epsilon(x_1)_{x_3} & \epsilon(x_2)_{x_3} & \epsilon(x_3)_{x_3} \end{bmatrix} = \begin{bmatrix} E_{11} & E_{21} & E_{31} \\ E_{12} & E_{22} & E_{32} \\ E_{13} & E_{23} & E_{33} \end{bmatrix} \begin{bmatrix} c_{11} & c_{21} & c_{31} \\ c_{12} & c_{22} & c_{32} \\ c_{13} & c_{23} & c_{33} \end{bmatrix}. \tag{40}$$

This formula, which may be demonstrated similarly for larger values of p, is preferable to (39) in that it presents the various values of a given x_i in the same column and utilizes the transpose of the inverse matrix.

Now if all ϵ_i and ϵ_{ij} have η as a bound, we may write

$$|E_{ij}| \leq \eta[1 + X_1 + X_2 + X_3]. \tag{41}$$

It follows that

$$E(x) \leq \eta[1 + \Sigma X] \begin{bmatrix} \pm 1 & \pm 1 & \pm 1 \\ \pm 1 & \pm 1 & \pm 1 \\ \pm 1 & \pm 1 & \pm 1 \end{bmatrix} \begin{bmatrix} c_{11} & c_{21} & c_{31} \\ c_{12} & c_{22} & c_{32} \\ c_{13} & c_{23} & c_{33} \end{bmatrix}. \tag{42}$$

The maximum value of $\epsilon(x_i)$ is obtained if the value of signs of the elements of the first row of the matrix containing unit elements is the same as the value of signs of the c_{1j}; the maximum value of $\epsilon(x_2)$ is obtained if the signs of the elements of the second row of the matrix

containing unit elements are the same as the signs of the c_{2j}, and the minimum value of $\epsilon(x_3)$ is obtained if the signs of the elements of the third row match those of the third column of c. Formal matrix multiplication then yields the value of $E(x)$.

If the adjugate, rather than the inverse, is available, we have the formula

$$(43)\quad E(x) = \frac{\eta[1 + \Sigma X]}{\Delta} \begin{bmatrix} \pm 1 & \pm 1 & \pm 1 \\ \pm 1 & \pm 1 & \pm 1 \\ \pm 1 & \pm 1 & \pm 1 \end{bmatrix} \begin{bmatrix} A'_{11} & A'_{12} & A'_{13} \\ A'_{21} & A'_{22} & A'_{23} \\ A'_{31} & A'_{32} & A'_{33} \end{bmatrix},$$

where the signs are determined in accordance with the foregoing instructions. This formula is also extendable to problems involving more variables.

An illustration of the problem of Table 17.7*g* is presented in Table 17.7*j*. The values of the A'_{ij} are formally premultiplied by the matrix

TABLE 17.7*j*

CALCULATION OF $\epsilon(x_i)$ WITH MATRIX MULTIPLICATION

	1	−2	3	7
	5	−4	−1	$\Delta = 8$
	1	4	3	0.875
	3	−4	1	$\eta = 0.01$
1 1 1	9	−4	3	0.00875
−1 1 −1	−7	12	3	
−1 1 1	−1	4	5	
	0.079	−0.035	0.026	
	−0.061	0.105	0.026	
	−0.009	0.035	0.044	

of unit terms after the appropriate signs are written. The results are multiplied by

$$\frac{\eta(1 + \Sigma X)}{\Delta} = \frac{(0.01)(7)}{8} = 0.00875.$$

Formal multiplication is not necessary since we need only to add the values in the x_i column, when multiplied by the sign of the corresponding value of the x_j column, to obtain the value of $k\epsilon(x_j)$. The matrix of unit terms does not need to appear. Thus the value of $\epsilon(x_3)_{x_2}$ in Table 17.7*j* is

$$\epsilon(x_3)_{x_2} = [(-1)(-1) + (3)(1) + (1)(-1)]0.00875 = 3\cdot(0.00875)$$
$$= 0.026.$$

An application to the illustration of Table 17.7*a* is made in Table 17.7*k*, and an application to the illustration of Table 17.7*b* in Table 17.7*l*.

TABLE 17.7*k*

$E(x)$ FOR ILLUSTRATION OF TABLE 17.7*a*

2	1	3	−2	9
66233	−16033	42069	−6503	$\Delta = 2305327$
56151	28558	33194	−52258	$3.904\cdot10^{-6}$
−53068	36236	18224	45899	$\eta = 0.005$
−35013	9659	−47056	53524	$1.952\cdot10^{-8}$
210465	−33370	104095	−158184	
−98163	90486	−37707	53668	
104329	39102	140543	−66386	
−210465	33370	−104095	158184	
0.0041	−0.0007	0.0020	−0.0031	
−0.0019	0.0018	−0.0007	0.0010	
0.0020	0.0008	0.0027	−0.0013	
−0.0041	0.0007	−0.0020	0.0031	

TABLE 17.7*l*

$E(x)$ FOR ILLUSTRATION OF TABLE 17.7*b*

−0.9366	0.0602	0.8152	1.1748	3.9868
2.0708	−0.1913	−0.7759	−1.0107	0.005
−0.1913	1.2842	−0.2185	−0.3552	0.019934
−0.7759	−0.2185	1.3988	0.2731	
−1.0107	−0.3552	0.2731	1.6941	
4.0487	−0.9018	−2.2293	−2.6227	
−0.4755	2.0492	−1.1145	−1.3117	
−3.6661	−1.6666	2.6663	3.3331	
−3.6661	−1.6666	2.6663	3.3331	
0.0807	−0.0180	−0.0444	−0.0523	
−0.0095	0.0408	−0.0222	−0.0261	
−0.0731	−0.0332	0.0532	0.0664	
−0.0731	−0.0332	0.0532	0.0664	

Comparison with the former results shows that these are much more general in that they exhibit the appropriate corrections to the other x_i as well as the correction to the x_i that is to be an extremum. This additional information can be obtained with a relatively small amount of work, as may be seen by comparison of these tables.

Other methods have been proposed for estimating bounds for $\epsilon(x_i)$. Milne [B], for example, proposes (6) rather than (7) as the basis of the calculation. Taking η as the bound of ϵ_i and ϵ_{ij}, we get

$$a\epsilon(x) \leq \eta(1 + \Sigma X_i)(1), \tag{44}$$

where (1) is a column matrix consisting of unit elements. The process consists in solving

$$a\epsilon'(x) \leq (1), \tag{45}$$

where all negative signs used in the calculation of the right side, in the direct and inverse solutions, are replaced by positive signs so as to keep the inequality true for all $\epsilon'(x)$. The values of $\epsilon'(x)$ are then multiplied by $k = \eta(1 + \Sigma X)$ to obtain the $\eta(x_i)$.

This solution has the desirable feature that its left part is identical with the solution of (2) so that the solution of (2) and (45) may be carried out simultaneously. This is illustrated in Table 17.7m, where the prob-

TABLE 17.7m

CALCULATION OF ERRORS, USING ABSOLUTE VALUES

x_1	x_2	x_3	a_{i4}	a'_{i4}	Sum	Error
2	1	−2	−6	−6	−5	1
1	1	1	2	2	5	1
−1	−2	3	12	12	12	1
2.0000	0.5000	−1.0000	−3.0000	−2.5000	−2.5000	0.5000
1.0000	0.5000	4.0000	10.0000	15.0000	15.0000	3.0000
−1.0000	−1.5000	8.0000	3.0000	4.0000	4.0000	0.7500
1	−2	3	7			
2	−1	4				
4.2500	6.0000	0.7500	$\eta = 0.01$			
0.298	0.420	0.053	$k = 0.07$			

lem of the illustration of Table 17.7*g* is solved by the compact method of single division. Here are included the approximate solution, the solution resulting from the row sum check, and the solution of errors, all of which have the same forward solution on the left.

The values of $\epsilon'(x)$ are multiplied by $k = 0.01[1 + 1 + 2 + 3] = 0.07$ to get 0.298, 0.420, 0.053. These bounds are larger than those determined in Table 17.7*g*, which are 0.079, 0.105, 0.044. Closer bounds for $\epsilon(x_1)$ and $\epsilon(x_2)$ could probably be obtained by this method if the back solution were not used and if the different bounds for $\epsilon(x_i)$ were obtained by different orders of elimination. If we are to do this much work, we might as well use the inverse or adjugate matrix and obtain results that not only are more precise, but also make possible the determination of the other $\epsilon(x_i)$ associated with some particular extreme value.

Willers [F] has presented another method that is proposed as a solution of (44). He finds $\eta(x)$, except that terms are always added on the right side (as in the method described by Milne), whereas on the left side the absolute values of the coefficients are taken and the products are subtracted. He suggests a separate elimination process for each $\eta(x_i)$. Again it seems more satisfactory to use the inverse or adjoint matrix, if we must do all the work demanded by this method, and thus obtain closer bounds. For purposes of illustration, the value of a bound for $\epsilon(x_3)$ for the problem just used is worked out in Table 17.7*n* by the method described by Willers, with the compact method of single division. In this case we find that the bound for $\epsilon(x_3)$ is 0.210. The

TABLE 17.7*n*

CALCULATION OF ERROR, USING WILLER'S METHOD

x_1	x_2	x_3	Error
2	1	2	0.07
1	1	1	0.07
1	2	3	0.07
2.0000	0.5000	1.0000	0.0350
1.0000	0.5000	0.0000	0.2100
1.0000	1.5000	2.0000	0.2100
		0.210	

method described by Milne gives 0.053, whereas the smallest (first-order) bound is actually 0.044.

17.8 Solutions of linear equations with some coefficients subject to error. Certain special cases are of sufficient importance to warrant special treatment. A less detailed presentation is given in the cases in which

(a) The $\epsilon(a)$ are 0.
(b) The $\epsilon(f)$ are 0.
(c) The $\epsilon(f)$ are 0 and all $\epsilon(a)$ are 0 except those of one column.
(d) The $\epsilon(f)$ are 0, and all $\epsilon(a)$ are 0 except those of one row.
(e) The $\epsilon(a)$ are 0 and so are all $\epsilon(f)$ except one element.
(f) The $\epsilon(f)$ are 0 and so are all $\epsilon(a)$ except one element.

(a) If the $\epsilon(a)$ are zero, (7.7.4) becomes

$$a\epsilon(x) = \epsilon(f)$$

so that we have

$$\epsilon(x) = c\epsilon(f) = \frac{A\epsilon(f)}{\Delta}. \tag{1}$$

(We should note that this formula is theoretically exact and is not a first-order approximation as are the formulas of section 17.7.) When $p = 3$, the formula (1) appears as

$$\begin{bmatrix} \epsilon(x_1) \\ \epsilon(x_2) \\ \epsilon(x_3) \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix} \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{bmatrix}. \tag{2}$$

Distinct values of ϵ_1, ϵ_2, ϵ_3 can be taken so as to maximize $\epsilon(x_1)$, or $\epsilon(x_2)$ or $\epsilon(x_3)$. In each case the corresponding values of the other $\epsilon(x_i)$ can be found by using the specified values ϵ_1, ϵ_2, ϵ_3 in (2). We can then write

$$\begin{bmatrix} \epsilon(x_1)_{x_1} & \epsilon(x_1)_{x_2} & \epsilon(x_1)_{x_3} \\ \epsilon(x_2)_{x_1} & \epsilon(x_2)_{x_2} & \epsilon(x_2)_{x_3} \\ \epsilon(x_3)_{x_1} & \epsilon(x_3)_{x_2} & \epsilon(x_3)_{x_3} \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix} \begin{bmatrix} \epsilon_{11} & \epsilon_{12} & \epsilon_{13} \\ \epsilon_{21} & \epsilon_{22} & \epsilon_{23} \\ \epsilon_{31} & \epsilon_{32} & \epsilon_{33} \end{bmatrix} \tag{3}$$

as in (17.7.39). Following the same development used in section 17.7, we get the same results, with the exception that the quantity $1 + \Sigma X_i$ is here replaced by 1.

The problems of Tables 17.7k and 17.7l are used as illustrations. It is understood now that the terms on the left are exact but that the f_i are subject to errors that are not greater than 0.005. The results are shown in Tables 17.8a and 17.8b.

TABLE 17.8a

ERRORS WITH $\epsilon(a) = 0$. USE OF ADJOINT

2	1	3	−2	1
66233	−16033	42069	−6503	$\Delta = 2305327$
56151	28558	33194	−52258	$4.337779 \cdot 10^{-7}$
−53068	36236	18224	45899	$\eta = 0.005$
−35013	9659	−47056	53524	$2.16889 \cdot 10^{-9}$
210465	−33370	104095	−158184	
−98163	90486	−37707	53668	
104329	39102	140543	−66386	
−210465	33370	−104095	158184	
0.0005	−0.0001	0.0002	−0.0003	
−0.0002	0.0002	−0.0001	0.0001	$E(x)$
0.0002	0.0001	0.0003	−0.0001	
−0.0005	0.0001	−0.0002	0.0003	

TABLE 17.8b

ERRORS WITH $\epsilon(a) = 0$. USE OF INVERSE

−0.9366	0.0602	0.8152	1.1748	1
2.0708	−0.1913	−0.7759	−1.0107	0.005
−0.1913	1.2842	−0.2185	−0.3552	0.005
−0.7759	−0.2185	1.3988	0.2731	
−1.0107	−0.3552	0.2731	1.6941	
4.0487	−0.9018	−2.2293	−2.6227	
−0.4755	2.0492	−1.1145	−1.3117	
−3.6661	−1.6666	2.6663	3.3331	
−3.6661	−1.6666	2.6663	3.3331	
0.0202	−0.0045	−0.0111	−0.0131	
−0.0024	0.0102	−0.0056	−0.0066	
−0.0183	−0.0083	0.0133	0.0167	
−0.0183	−0.0083	0.0133	0.0167	

TABLE 17.8*c*

ERRORS WITH $\epsilon(f) = 0$. USE OF ADJUGATE

2	1	3	−2	8
66233	−16033	42069	−6503	$\Delta = 2305327$
56151	28558	33194	−52258	$347022 \cdot 10^{-6}$
−53068	36236	18224	45899	$\eta = 0.005$
−35013	9659	−47056	53524	$173511 \cdot 10^{-8}$
210465	−33370	104095	−158184	
−98163	90486	−37707	53668	
104329	39102	140543	−66386	
−210465	33370	−104095	158184	
0.0037	−0.0006	0.0018	−0.0027	
−0.0017	0.0016	−0.0006	0.0009	
0.0018	0.0007	0.0024	−0.0012	
−0.0037	0.0006	−0.0018	0.0027	

TABLE 17.8*d*

ERRORS WITH $\epsilon(f) = 0$. USE OF INVERSE

−0.9366	0.0602	0.8152	1.1748	2.9868
2.0708	−0.1913	−0.7759	−1.0107	0.005
−0.1913	1.2842	−0.2185	−0.3552	0.014034
−0.7759	−0.2185	1.3988	0.2731	
−1.0107	−0.3552	0.2731	1.6941	
4.0487	−0.9018	−2.2293	−2.6227	
−0.4755	2.0492	−1.1145	−1.3117	
−3.6661	−1.6666	2.6663	3.3331	
−3.6661	−1.6666	2.6663	3.3331	
0.0605	−0.0135	−0.0333	−0.0392	
−0.0071	0.0306	−0.0166	−0.0196	
−0.0547	−0.0249	0.0398	0.0498	
−0.0547	−0.0249	0.0398	0.0498	

(*b*) If the $\epsilon(f)$ are zero, (17.7.7) becomes

$$\epsilon(x) = -c\epsilon(a)x. \tag{4}$$

This, like (17.7.7), is a first-order approximation formula. A treatment of (4) parallels the treatment (17.7.7) shown in (31) to (43) of the last section, and we arrive at the same results with ΣX replacing $1 + \Sigma X$.

Illustrations are given in Tables 17.8*c* and 17.8*d*. We now know that

TABLE 17.8*c*

ERRORS FOR $\epsilon_{ik} \neq 0$. USE OF ADJOINT

2	1	3	−2	3
66233	−16033	42069	−6503	$\Delta = 2305327$
56151	28558	33194	−52258	$1301334 \cdot 10^{-7}$
−53068	36236	18224	45899	$\eta = 0.005$
−35013	9659	−47056	53524	$6.506670 \cdot 10^{-9}$
210465	−33370	104095	−158184	
−98163	90486	−37707	53668	
104329	39102	140543	−66386	
−210465	33370	−104095	158184	
0.0014	−0.0002	0.0007	−0.0010	
−0.0006	0.0006	−0.0002	0.0003	
0.0007	0.0003	0.0009	−0.0004	
−0.0014	0.0002	−0.0007	0.0010	

the ΣX in the formula is due to the errors of the coefficients on the left and that the 1 in the formula is due to the error of the coefficients on the right.

(*c*) A special case of (*b*) is that one in which all $\epsilon(a_{ij})$ are zero except those of some column, k, for instance. In this case $|\epsilon(a_{ij})| \leq \eta$. The matrix equation (4) becomes, when $p = 2$ and $k = 2$,

$$\begin{bmatrix} \epsilon(x_1) \\ \epsilon(x_2) \\ \epsilon(x_3) \end{bmatrix} = -\begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix} \begin{bmatrix} 0 & \epsilon_{12} & 0 \\ 0 & \epsilon_{22} & 0 \\ 0 & \epsilon_{32} & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}. \tag{5}$$

Expansion of (5) shows that x_2 is a factor of every $\epsilon(x_i)$. Furthermore, the values of ϵ_{12}, ϵ_{22}, ϵ_{32} differ, depending on whether it is $\epsilon(x_1)$, $\epsilon(x_2)$, or $\epsilon(x_3)$ that is to be maximized. The general argument also follows

that of the last section, with the value of X_k replacing the value $1 + \Sigma X_i$. This result is of some import since it enables us to break down the total maximum error into components that are due to the different columns of the problem.

TABLE 17.8*f*

ERRORS FOR $\epsilon_{ik} \neq 0$. USE OF INVERSE

−0.9366	0.0602	0.8152	1.1748	0.8152
2.0708	−0.1913	−0.7759	−1.0107	0.005
−0.1913	1.2842	−0.2185	−0.3552	0.004076
−0.7759	−0.2185	1.3988	0.2731	
−1.0107	−0.3552	0.2731	1.6941	
4.0487	−0.9018	−2.2293	−2.6227	
−0.4755	2.0492	−1.1145	−1.3117	
−3.6661	−1.6666	2.6663	3.3331	
−3.6661	−1.6666	2.6663	3.3331	
0.0165	−0.0037	−0.0091	−0.0107	
−0.0019	0.0084	−0.0045	−0.0053	
−0.0149	−0.0068	0.0109	0.0136	
−0.0149	−0.0068	0.0109	0.0136	

Illustrations are given in Tables 7.8*e* and 7.8*f*. In each case the coefficients of the third column of the original equations are approximate numbers, $\eta = 0.005$, whereas the others are exact.

(*d*) A somewhat different result appears when the coefficients of some one row, for example, a_{kj}, only are subject to error. In this case (4) becomes, when $p = 3$,

$$(6)\quad \begin{bmatrix} \epsilon(x_1) \\ \epsilon(x_2) \\ \epsilon(x_3) \end{bmatrix} = -\begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ \epsilon_{21} & \epsilon_{22} & \epsilon_{23} \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

$$= -\begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix} \begin{bmatrix} 0 \\ \epsilon_{21}x_1 + \epsilon_{22}x_2 + \epsilon_{23}x_3 \\ 0 \end{bmatrix}$$

$$= -\begin{bmatrix} c_{12}(\epsilon_{21}x_1 + \epsilon_{22}x_2 + \epsilon_{23}x_3) \\ c_{22}(\epsilon_{21}x_1 + \epsilon_{22}x_2 + \epsilon_{23}x_3) \\ c_{32}(\epsilon_{21}x_1 + \epsilon_{22}x + \epsilon_{23}x_3) \end{bmatrix}.$$

In general, it can be shown that

$$\epsilon(x_i) = -c_{ik}(\epsilon_{k1}x_1 + \epsilon_{k2}x_2 + \epsilon_{k3}x_3 + \cdots + \epsilon_{kp}x_p). \tag{7}$$

This has a maximum value when the signs of ϵ_{kj} are taken opposite to those of x_j, the value of $\Sigma\epsilon_{kj}x_j$ is found and the sign taken opposite to that of c_{ik}. If the ϵ_{kj} have a bound, η, the various values of $\epsilon(x_i)$ for the maximum of a specific $\epsilon(x_i)$ are proportional to the values of c_{ik}. The ϵ_{kj} that give a maximum for any specific $\epsilon(x_i)$ give either a minimum or a maximum for each other x_i. Only one row of values is necessary. This is obtained by multiplying the values of c_{ik} by $\pm\eta\Sigma X_i$.

Illustrations are presented in Tables 17.8*g* and 17.8*h*, where the third

TABLE 17.8*g*

ERRORS FOR $\epsilon_{kj} \neq 0$. USE OF ADJOINT

2	1	3	−2	8	$3.47022 \cdot 10^{-6}$
−53068	36236	18224	45899	2305327	$\eta = 0.005$
±0.0009	±0.0006	±0.0003	±0.0008		$1.73511 \cdot 10^{-8}$

TABLE 17.8*h*

ERRORS FOR $\epsilon_{kj} \neq 0$. USE OF INVERSE

−0.9366	0.0602	0.8152	1.1748	2.9868
−0.7759	−0.2185	1.3988	0.2731	0.005
±0.0116	±0.0033	±0.0209	±0.0041	0.014934

rows of the original equations have errors of not more than 0.005. The third rows of the transpose of the adjugate and inverse are used.

The problem might arise in which the whole row, including the term on the right, might be subject to error. The foregoing technique could be used with $1 + \Sigma X_i$ replacing ΣX_i.

(*e*) In case all errors are zero, excepting one in the kth row of f, we have the technique of (*d*) with the value ΣX_i being replaced by 1. This is a special case of (*a*) so that the results are theoretically exact and are not first-order approximations.

(*f*) The case in which all $\epsilon(f) = 0$ and all $\epsilon(a) = 0$, except the element in the kth row and lth column, is a special case of (*c*) and (*d*). The basic matrix equation is (17.7.7), with the $\epsilon(a)$ identically zero except for the element e_{kl}. The expansion shows, when $p = 3$, $k = 2$, and $e = 3$,

$$(8) \qquad \begin{bmatrix} \epsilon(x_1) \\ \epsilon(x_2) \\ \epsilon(x_3) \end{bmatrix} = - \begin{bmatrix} c_{12} & \epsilon_{23} & x_3 \\ c_{22} & \epsilon_{23} & x_3 \\ c_{32} & \epsilon_{23} & x_3 \end{bmatrix} = - \frac{1}{\Delta} \begin{bmatrix} A_{12} & \epsilon_{23} & x_3 \\ A_{22} & \epsilon_{23} & x_3 \\ A_{32} & \epsilon_{23} & x_3 \end{bmatrix}.$$

In general the formula is

$$(9) \qquad \epsilon(x_i) = -c_{ik}\epsilon_{kl}x_l.$$

If $|\epsilon_{kl}| \leq \eta$, the values of $\epsilon(x_i)$ for the maximum value of a specific x_i are proportional to the values of c_{ik}. The result is just like that of (*d*), except that x_l is used rather than ΣX_i.

This method can be used to correct the answers to a set of exact equations when it is learned that one of the coefficients has an error if the error is known and is relatively small. Suppose, for example, that a_{13} in Table 5.6*a* is 15.01 rather than 15. Then ϵ_{13} is 0.01, and the first-order corrections to the solutions are

$$(10) \quad \epsilon(x_i) = \frac{-A_{i1}\epsilon_{13}x_3}{\Delta} = \frac{-A_{i1}(0.01)(3)}{2{,}305{,}327} = -A_{i1}(1.301334 \cdot 10^{-8}).$$

Now from Table 14.5*e* we see that the values of A_{i1} are 66233, −16033, 42069, −6503 so that the corrections are −0.00086, 0.00021, −0.00055, 0.00008, and the new values of x_i are 1.99914, 1.00021, 2.99945, −1.99992. When substituted in the new equations, these give results that agree with the right side to five significant places.

If we change the $a_{13} = 15$ to $a_{13} = 16$, ϵ_{13} is 1 and the errors are −0.086, 0.021, −0.055, and 0.008 so that the corrected answers are 1.914, 1.021, 2.945, and −1.992. Although the error is somewhat large for the use of a first-order approximation, a substitution in the new equations gives results that agree with the right side to three significant places. In any case the measure of the error in the f that results from the first-order approximation may be calculated from the neglected term $\epsilon(a)\epsilon(x)$. This becomes, when $\epsilon_{13} = 1$ with all other $\epsilon_{ij} = 0$, the value of $\epsilon(x)$ itself. In this case the actual figures on the right side and these resulting from a final verification must agree to one decimal place if all calculations are exact.

17.9 The errors of the inverse and the adjoint. We can obtain formulas for the errors of the terms of the inverse and adjoint as special

cases of the formulas of section 17.7 since f may be a square matrix and is not necessarily a column vector. If $f = I$ with $\epsilon(f) = 0$, (17.7.2) becomes

$$ax = I \qquad \text{with} \qquad x = a^{-1} = c \tag{1}$$

and (17.7.7) becomes

$$\begin{aligned} \epsilon(x) = \epsilon(a^{-1}) &= -a^{-1}\epsilon(a)a^{-1} \\ &= \epsilon(c) = -c\epsilon(a)c. \end{aligned} \tag{2}$$

This formula is a matrix generalization of the calculus formula for $d(x^{-1})$ [G]. The matrices c, $\epsilon(a)$, and $\epsilon(c)$ are now all p by p matrices. When $p = 3$, we have

$$\begin{aligned} &\begin{bmatrix} \epsilon(c_{11}) & \epsilon(c_{12}) & \epsilon(c_{13}) \\ \epsilon(c_{21}) & \epsilon(c_{22}) & \epsilon(c_{23}) \\ \epsilon(c_{31}) & \epsilon(c_{32}) & \epsilon(c_{33}) \end{bmatrix} \\ &= -\begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix} \begin{bmatrix} \epsilon_{11} & \epsilon_{12} & \epsilon_{13} \\ \epsilon_{21} & \epsilon_{22} & \epsilon_{23} \\ \epsilon_{31} & \epsilon_{32} & \epsilon_{33} \end{bmatrix} \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix}. \end{aligned} \tag{3}$$

Furthermore the $p^2 = 9$ of the ϵ_{ij} can be assigned in $p^2 = 9$ ways so as to maximize the $p^2 = 9$ values of $\epsilon(c_{ij})$. Each of these nine sets of ϵ_{ij} can in turn be inserted in (3) to obtain the $p^2 = 9$ different error matrices. Such extensive treatment is not appropriate to this book, but a method that arrives at the maximum value of each $\epsilon(c_{ij})$ should be presented. The argument, which is similar to the arguments of the last two sections, is presented in rather brief form.

Now any specific $\epsilon(c_{kl})$ can be written in the form

$$\epsilon(c_{kl}) = -[c_{k1} \quad c_{k2} \quad c_{k3}] \begin{bmatrix} \epsilon_{11} & \epsilon_{12} & \epsilon_{13} \\ \epsilon_{21} & \epsilon_{22} & \epsilon_{23} \\ \epsilon_{31} & \epsilon_{32} & \epsilon_{33} \end{bmatrix} \begin{bmatrix} c_{1l} \\ c_{2l} \\ c_{3l} \end{bmatrix}. \tag{4}$$

Formal expansion shows that

$$\epsilon(c_{kl}) = -\sum_{i,j} c_{ki}\epsilon_{ij}c_{jl}. \tag{5}$$

This formula holds for the general case as can be seen from (2). The maximum absolute value of $\epsilon(c_{kl})$ is obtained by taking ϵ_{ij} opposite in sign to $c_{ki}c_{jl}$. (We may calculate the signs of each ϵ_{ij} from the inverse matrix and attach the errors ϵ_{ij} to the a_{ij} and then calculate the inverse matrix having the maximum $\epsilon(c_{ij})$ if we wish.) Instead we shall calcu-

late maximum $\epsilon(c_{kl})$ from (5). If $|\epsilon_{ij}| \leq \eta$, we have

$$|\epsilon(c_{kl})| \leq \eta \sum_{ij} |c_{ki}||c_{jl}| \tag{6}$$

and thence

$$|\eta(c_{kl})| \leq \eta \left(\sum_i |c_{ki}|\right)\left(\sum_j |c_{jl}|\right). \tag{7}$$

We need only sum the absolute values in the kth row and the lth column of the inverse matrix to obtain the product and multiply by η to obtain a bound for the maximum error of ϵ_{kl}. If we use the transpose of the inverse, according to our custom, we must use the elements of the kth column and those of the lth row to obtain the maximum error for c_{kl}. But this should be placed in the l,k position of the transpose of the matrix of corrections. We can accomplish this easily if we write the sums at the bottom of each row and the right of each column, then place the product of the sums times η at the intersection of the row and column. These are the error parts of approximation-error numbers so that the notation of Chapter 2 may be used.

The process is illustrated in Table 17.9*a*, where the problem of the earlier sections is used with $\eta = 0.005$. The maximum errors for the

TABLE 17.9*a*

THE INVERSE WITH MAXIMUM ERRORS—ILLUSTRATION OF TABLE 17.7*b*

−0.9366(807)	0.0602(408)	0.8152(532)	1.1748(664)	3.9868
2.0708(820)	−0.1913(415)	−0.7759(540)	−1.0107(675)	4.0487
−0.1913(415)	1.2842(210)	−0.2185(273)	−0.3552(342)	2.0492
−0.7759(540)	−0.2185(273)	1.3988(355)	0.2731(444)	2.6663
−1.0107(675)	−0.3552(342)	0.2731(444)	1.6941(555)	3.3331
4.0487	2.0492	2.6663	3.3331	0.0015

$\Delta = 0.3660(221)$

values of x_i (see Table 17.7*b*) are calculated similarly and placed in the top row. If the original a_{ij} were correct to four decimal places, then the value of η is 0.00005 and the maximum errors are 1/100 of the errors. In this case the approximation-error numbers of the solution and the inverse matrix have small errors, as is shown in Table 17.9*b*. If a is symmetric, c is symmetric, and we see from (7) that $\eta(c_{kl}) = \eta(c_{lk})$ so that symmetry may be used in eliminating $\dfrac{p(p-1)}{2}$ of the calculations.

TABLE 17.9*b*

RESULTS OF TABLE 17.9*a* WITH $\eta = 0.00005$

−0.9366(8)	0.0602(4)	0.8152(5)	1.1748(7)
2.0708(8)	−0.1913(4)	−0.7759(5)	−1.0107(7)
−0.1913(4)	1.2842(2)	−0.2185(3)	−0.3552(3)
−0.7759(5)	−0.2185(3)	1.3988(4)	0.2731(4)
−1.0107(7)	−0.3552(3)	0.2731(4)	1.6941(6)

For work with the adjoint, we have from (7)

$$\eta(A_{kl}) \leq \frac{\eta}{\Delta} \sum_i |A_{ki}| \sum_j |A_{jl}|. \tag{8}$$

The method is similar to that of Table 17.9*a*, but a final division by Δ is required.

The method is applied to the illustration of Table 17.7*k* in Table 17.9*c*. The values of x_i are presented in the first row. All the results

TABLE 17.9*c*

THE ADJUGATE WITH MAXIMUM ERRORS—ILLUSTRATION OF TABLE 17.7*k*

2.0000(41)	1.0000(18)	3.0000(27)	−2.0000(31)	9
66233(60)	−16033(26)	42069(40)	−6503(45)	130838
56151(78)	28558(33)	33194(52)	−52258(58)	170161
−53068(70)	36236(30)	18224(47)	45899(53)	153427
−35013(66)	9659(29)	−47056(44)	53524(50)	145252
210465	90486	140543	158184	2305327 (2998)

$\eta = 0.005$

are given in the form of approximation-error numbers.

The value of $\eta(\Delta)$ of (17.6.1) may be computed along with the other errors. We need only to add the values of the last row (or column) and multiply by 0.005 to obtain

$$\eta(\Delta) = (599678)(0.005) = 2998.$$

This value of $\eta(\Delta)$ is added to that of Δ in the lower right corner.

Similarly, the values of $\eta(\Delta)$ may be computed along with other errors where the inverse is used. The value of the determinant is calculated, with any variant of the method of single division, by using (10.6.6). From Table 6.4*a* we know that the approximate value of Δ is

$$\Delta = (1.0000)(0.8400)(0.7381)(0.5903) = 0.3660$$

whereas application of (17.6.2) to Table 17.9*a* gives

$$\eta(\Delta) = (0.3660)(0.005)[4.0487 + 2.0492 + 2.6663 + 3.3331]$$
$$= (0.3660)(0.005)(12.0973) = 0.0221.$$

The value of $\Delta = 0.3660(221)$ is placed in the last line of Table 17.9*a*. We can transform the approximation-error value of the adjoint to approximation-error values of the inverse by the division by the approximation-error value of Δ to obtain approximation-error values of the adjoint by the methods of Chapter 2.

17.10 The errors of determinantal ratios. Etherington [D] gives general first- and second-order formulas for the errors in the quotient of certain determinants together with specific formulas for first-order terms that are similar to some of the formulas of the sections above. There is a quite general treatment of the first-order approximations to determinantal ratios in the above sections since the X_i of section 17.7 and the c_{kl} of section 17.8 are determinantal ratios. Etherington also gives a probability approach to the variation in the errors of the X_i, which he calls a quotient of determinants. This type of approach is not presented here since the actual range of errors is more satisfactory than a standard deviation of errors and since the standard deviation of errors, under suitable assumptions regarding the distribution of errors, is not much smaller than the semi-range that we are using.

17.11 Error control. The errors discussed in the preceding sections are type *a* errors, the inherent errors, which are present in the linear problem itself. We have assumed in the discussion that the rounding-off errors, the type *b* errors, were non-existent or at least trivial. A good way to avoid type *b* errors, if it is practical, is to provide an exact solution of (17.7.2). This was done in the problem of Tables 5.6*a* and 14.5*e*, a problem that has been used extensively for illustration so there are no type *b* errors in the solution.

Because exact methods are not always practical approximate methods are used. In this case, however, we can carry the incomplete numbers of the solutions of (17.7.2) to more places than are indicated by the explicit values of (17.7.1). This is done in the other illustration, which is used extensively in this chapter because the solutions of (17.7.2) are

carried to four decimal places while the equations (17.7.1) are accurate to two decimal places. In this case, presumably, the amount of the accumulation of the type b errors is trivial with respect to the amount of accumulation of the type a errors, so that the type b errors are neglected in solving the errors of the results.

Frequently the incomplete numbers of the solution are carried to the number of places indicated by the type a errors; therefore, the original type a errors and type b errors are of the same size, that is, the rounding-off errors of a given decimal position. In this case it is customary to speak of all errors as rounding-off errors, that is, type b errors. Thus the values of the identity matrix of Table 13.6a, though they are known to be exactly unity and zero, are recorded as four-decimal-place approximations to unity and zero for use with four-decimal-place incomplete numbers. The four-decimal incomplete numbers on the left, though they result from type a errors, have the form of the rounding-off errors. Previously some writers have included them in their discussions of rounding-off errors; the rounding-off errors of these writers, therefore, included type a errors.

Now, under certain conditions, the rounding-off errors may cumulate in such a way as to invalidate the assumption that the accumulation of them is small. Hotelling [H.1], for example, indicates by continued application of the type of inequalities used in Chapter 2, that, if $p = 11$, incomplete numbers must be carried to at least seven decimal places if we are to be sure that the solutions of the equations (involving correlations) are correct to one decimal place. Of course the specification of the type of matrix may enable us to cut down the size of the bounds, as von Neumann and Goldstine [A] and Satterthwaite [C] have done. Turing also [H.2] has given an extensive discussion. A detailed treatment of this subject is hardly appropriate to this book, but a statement of some useful results and methods of Satterthwaite is in order.

The norm of a matrix, used by Hotelling and Satterthwaite, is defined as

$$N(a) = \sqrt{\sum_i \sum_j a_{ij}^2}. \tag{1}$$

They give various inequalities involving norms. Satterthwaite made an analysis, using norms, of maximum accumulation of rounding-off errors resulting from using the method of single division. He found that if the norm of the matrix $(a - I)$ is less than 0.35, and if the number of the equations is less than 21, the accumulation of the errors in the results does not involve more than the last three decimal positions of the incomplete numbers of the results. We wish to apply this to a matrix

or set of equations in which the a_{ij} are recorded to two decimal places. If a is non-singular and if the norm of $(a - I) \leq 0.35$, we may carry the incomplete numbers through the forward and back solutions to five decimal positions, with the knowledge that the accumulation of rounding-off errors does not change the second decimal position of the solutions of the equations.

We wish to solve the equations of (2) and to get the inverse of the coefficients on the left of (2) with this theory. The equations (2) are the equations of Table 6.4a written to two decimal positions:

$$
\begin{aligned}
1.00x_1 + 0.40x_2 + 0.50x_3 + 0.60x_4 &= 0.20 \\
0.40x_1 + 1.00x_2 + 0.30x_3 + 0.40x_4 &= 0.40 \\
0.50x_1 + 0.30x_2 + 1.00x_3 + 0.20x_4 &= 0.60 \\
0.60x_1 + 0.40x_2 + 0.20x_3 + 1.00x_4 &= 0.80.
\end{aligned} \tag{2}
$$

Calculation shows that $N(A - I) = \sqrt{2.12} = 1.456$. This is considerably in excess of the value of 0.35 specified by Satterthwaite.

We can multiply each side of (2), however, by some exact matrix that will not change the values of x_i. If we multiply by c_1, an approximation to the inverse c, c_1a is an approximation to the identity matrix, and $N(a - I)$ is small. We can make the calculation to see whether or not $N(a - I) \leq 0.35$.

The calculational work is presented in Table 17.11a. An approximation to the inverse of a is obtained and its transpose is written at the left

TABLE 17.11a

ERROR CONTROL WITH THE METHOD OF SINGLE DIVISION—SOLUTION OF EQUATIONS

2.1	−0.2	−0.8	−1.0	1.00	0.40	0.50	0.60	0.20
−0.2	1.3	−0.2	−0.4	0.40	1.00	0.30	0.40	0.40
−0.8	−0.2	1.4	0.3	0.50	0.30	1.00	0.20	0.60
−1.0	−0.4	0.3	1.7	0.60	0.40	0.20	1.00	0.80
1				1.02	0.00	−0.01	0.02	−0.94
	1			−0.02	1.00	0.01	−0.04	0.04
		1		0.00	0.02	1.00	0.02	0.84
			1	0.01	−0.03	0.02	1.00	1.18
				1.02000	0.00000	−0.00980	0.01961	−0.92157
				−0.02000	1.00000	0.00980	−0.03961	0.02157
				0.00000	0.02000	0.99980	0.02080	0.83974
				0.01000	−0.03000	0.02039	0.99859	1.17486
				0.91825	0.06012	0.81530	1.17486	

of the table. Column-by-column multiplication of c'_1 with a and f results in the matrix equation

$$c_1 a x = c_1 f \tag{3}$$

whose synthetic form occupies the second four rows of the table. The solution of (3) is identical with the solution of (2) if the multiplication by c_1 is exact. The norm of $c_1 a - I$ is $\sqrt{0.0052} = 0.072$, and this is considerably lower than the required value of 0.35. The compact method of single division is used in solving (3). The theory calls for five decimal places in the solution if the results are to be accurate to at least two decimal places. Comparison with a solution accurate to more places (see section 4.8) shows that the errors actually appear first in the fourth decimal position.

A similar technique can be used in obtaining the inverse. Post-multiplication of $ax = I$ by c_1 gives

$$c_1 a x = c_1. \tag{4}$$

Satterthwaite indicates that the norm of c_1 should be less than 1. If the norm of c_1 is not less than 1, a power of ten can be factored from c_1. Now the norm of c_1 in Table 17.11a is not less than 1, but the norm of $1/(10c_1)$ is less than 1. We therefore solve the equation

$$c_1 \frac{ax}{10} = \frac{c_1}{10} \tag{5}$$

and multiply the results $x_i/10$ by 10 to obtain the value of a^{-1}. This is done by the compact method of single division, in Table 17.11b,

TABLE 17.11b

ERROR CONTROL WITH THE METHOD OF SINGLE DIVISION—CALCULATION OF THE INVERSE

1.02	0.00	−0.01	0.02	0.21	−0.02	−0.08	−0.10
−0.02	1.00	0.01	−0.04	−0.02	0.13	−0.02	−0.04
0.00	0.02	1.00	0.02	−0.08	−0.02	0.14	0.03
0.01	−0.03	0.02	1.00	−0.10	−0.04	0.03	0.17
1.02000	0.00000	−0.00980	0.01961	0.20588	−0.01961	−0.07843	−0.09804
−0.02000	1.00000	0.00980	−0.03961	−0.01588	0.12961	−0.02157	−0.04196
0.00000	0.02000	0.99980	0.02080	−0.07970	−0.02260	0.14046	0.03085
0.01000	−0.03000	0.02039	0.99856	−0.10106	−0.03551	0.02731	0.16934
0.20710	−0.01912	−0.07760	−0.10106				
0.01913	0.12842	−0.02186	−0.03551				
0.07759	−0.02186	0.13989	0.02731				
−0.10109	−0.03552	0.02733	0.16934				

using five-decimal-place incomplete numbers. The back solution results in $a^{-1}/10$ so that the value of a^{-1} is obtained by moving the decimal point one position to the left. This process guarantees that $a^{-1}/10$ is correct to two decimal places and hence that a^{-1} is correct to one decimal place. Actually the errors do not occur before the fourth decimal position of a^{-1}.

17.12 The solution of adjusted equations. Sets of equations obtained by adjusting certain of the coefficients of certain equations previously solved might be called *adjusted equations*. Can we obtain the solution of the adjusted equations from the solution of the original set? For example, a mistake has been found in one of the coefficients that was used in obtaining the solution. Can we adjust the solution for this known change in the coefficient? Can we solve the problem if a whole row (or a whole column) of coefficients is adjusted?

The formulas and methods of section 17.8 may be used to obtain fairly satisfactory answers to these problems. If the adjusted terms are on the right side of the equations, these methods give exact answers to the questions. They give satisfactory answers also if the adjustments are relatively small, when the adjusted terms are on the left side, since the restriction to first-order error terms is not important. They are not very satisfactory when the adjustment is large as, for instance, when the sign of a given coefficient is found to be in error. Some more specific methods should be available.

The general question includes the linear problems of the solution of simultaneous equations and the calculation of the inverse matrix. But the first of these can be solved if we know the answers to the second, since $x = a^{-1}f = cf$. We concentrate on the problem of finding the adjusted values of c.

This problem has received the attention of Sherman and Morrison [I], who have worked out determinantal formulas (a) for the case in which a single element is adjusted and (b) for the case in which the elements of a row or column are adjusted.

These determinantal formulas are not given here. Instead a matrix proof is presented that leads to an equivalent computational technique.

The proof is given for the general case (b) since this covers the special case (a). We first assume that the adjusted elements are in the last column of a, where a is non-singular. Let ϵ be the matrix having the same number of rows and columns of a that contains the adjustments for a. In the case under consideration, ϵ is identically zero except for the elements of the last column, which are the adjustments. We wish to find the value of $(a + \epsilon)^{-1}$ when $a + \epsilon$, of course, must be non-singular. Now

$$(1) \qquad (a+\epsilon)^{-1} = [a(I + a^{-1}\epsilon)]^{-1} = (I + a^{-1}\epsilon)^{-1}a^{-1}$$

and, if $a^{-1} = c$, we have

$$(2) \qquad (a+\epsilon)^{-1} = (I + c\epsilon)^{-1}c.$$

Expansion of $(I + c\epsilon)$ gives

$$(3) \qquad (I + c\epsilon) = \begin{bmatrix} I_{p+1} & E \\ 0 & s \end{bmatrix},$$

where I_{p-1} is a $p-1$ by $p-1$ identity matrix, E is a column vector with the $p-1$ components $E_i = \sum_j c_{pj}\epsilon_{jp}$ and

$$s = 1 + E_p = 1 + \sum_j c_{pj}\epsilon_{jp}.$$

Application of (16.2.9) gives

$$(4) \qquad (I + c\epsilon)^{-1} = \begin{bmatrix} I_{p-1} & -\dfrac{E}{s} \\ 0 & \dfrac{1}{s} \end{bmatrix}.$$

The matrix $(I + c\epsilon)^{-1}$ is then an identity matrix bordered with values $-E/s$, $1/s$ and zeros. The restriction to the last column may now

TABLE 17.12*a*

ADJUSTMENT OF COLUMNS—MODIFIED SHERMAN-MORRISON TECHNIQUE

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	0.9136	0.6773	0.7578		0.5823		0.8268		0.8981
	1.2652	1.9762	2.1252		1.5726		1.7783		0.0046
	1.0000	0.0663	0.0720		0.0521		0.0652		0.0047
	0	0.0548	0.0582		0.0698		0.0472		1.0000
	0	1.0000	1.0000		1.0000		1.0000		0.0013
					1.06393		0.02334		0.00503
1	−0.03142		−0.02785						
−0.00069	−1.86114	1.27928	2.48790	−0.45976	−1.44783	0.18047	1.67058	0.01167	−2.86993
	−3.33087		−0.38715	1	4.10223		2.97107		3.83494
	5.19189		−1.65074		−2.65480	1	−4.64284		0.03513
	0.09576		−0.00845		−0.07679		0.91454	1	−0.10231
adjust	−0.0805		−0.1490		−0.0057		−0.0084		0.0000
	0.00054		0.78169		0.35939		−0.14107		−0.00912
	−0.03014		−0.02957		1.06493		0.02219		0.00701
	−2.38092		3.18272		−1.85154		2.13714		−3.67144
	−2.47519		−1.98099		4.76765		2.20300		5.15442
	4.85601		−1.20175		−2.91600		−4.34135		−0.48281
	0.07404		0.02058		−0.09368		0.03404		−0.13580

be removed. The value of $(I + c\epsilon)^{-1}$ when ϵ_k are elements in the kth column is an identity matrix, with the exception of the kth column, which is similar to the last column of (4).

The transpose of ϵ may be used in forming a row-by-row multiplication of c and ϵ' to get $c\epsilon$, whereas the transpose of $(I + c\epsilon)^{-1}$ can be used in obtaining (2) by a column-by-column multiplication of the transpose $(I + c\epsilon)^{-1}$ and c. This is illustrated in Table 17.12a, where the problem proposed by Sherman and Morrison is treated by this modification of their technique. They gave the matrix a and the matrix c with new values to replace the values of the second column of a. The values of a and c are placed in the even-numbered columns of Table 17.12a; the new values are placed in column (3) beside the values in column (4) that they replace.

The adjustments to the a matrix are obtained by subtracting the old values from the new values. These are placed in the first row below the matrix c. Row-by-row multiplication then results in the values of E, which are placed in the next row. The value of s is found in the second column and results from the operation

$$\begin{aligned} s &= 1 + (-1.86114)(-0.0805) + (2.48790)(-0.1490) \\ &\quad + (-1.44733)(-0.0057) \\ &\quad + (1.67058)(-0.0034) + (-2.86993)(0.0000) \\ &= 0.78169. \end{aligned}$$

The value of $1/s$ is then placed in column 3 parallel to the c_{22} term, and the values of E/s are placed in the second row in the odd-numbered columns. Unity is placed in the diagonal terms, and the column-by-column multiplication is accomplished. The result is presented in the last five rows.

The same technique can be used when a single element is to be adjusted. In this case all the values of ϵ, except the one, are zero.

A technique for the adjustments in a given row can be based on the matrix formula

$$(5)\quad (a + \epsilon)^{-1} = (\epsilon + a)^{-1} = [(\epsilon a^{-1} + I)a]^{-1} = a^{-1}(\epsilon a^{-1} + I)^{-1}$$
$$= c(\epsilon c + I)^{-1}$$

or we may simply use the transpose of a and a^{-1}. The rows are transposed to columns so that we may use the method above and take the transpose of the result.

17.13 The deletion of variables. We sometimes find that a given variable makes little contribution to the solution so that we may wish to delete it from the solution and from the inverse or adjoint matrix. We can do this easily if the variable to be deleted is the last variable treated in the elimination process since the last terms in the forward solution may then be ignored. We have more difficulty when the deleted variable does not occupy this position. We do have methods for determining the values of the deleted adjoint or inverse from the values of the adjoint or inverse of the original matrix. The value of the deleted solution may then be found by applying

$$x = cf = \frac{Af}{\Delta}, \tag{1}$$

where c and A are the deleted results and f is the matrix on the right that comes from the deletion of one equation.

Consider the system (4.1.1) and delete variable x_k (column k). Then the system has one more equation than it has unknowns so that we may delete a row. We can delete any row, but symmetry is maintained, if it is originally present, if row k is deleted. In the following discussion we assume that the kth variable and the kth row are deleted.

We first derive a technique for computing the adjugate that results from the deletion of the kth row and the kth column of a, after which application is made to the effect on the inverse and on the solution of the equations.

The present problem is, in a sense, the opposite of the problem of Tables 16.2*c* and 16.2*d*. Then (16.2.15) may be written in the form

$$A'_{ij)k(} = \frac{A'_{kk}A'_{ij} - A'_{ik}A'_{kj}}{\Delta}, \tag{2}$$

where $A'_{ij)k(}$ is the element in the ij position in that adjugate that results from the elimination of row k and column k from a.

The use of (1) demands the determinant of a, as well as the elements of the adjugate. This value is usually known if the value of the adjugate is known. If not, it can be found by multiplication of any row of A' by the corresponding row of a.

We may wish to eliminate a second variable after the first variable has been eliminated. Formula (1) is also applicable, although the terms must be redefined. For instance, the new Δ is the old A'_{kk}. The resulting technique, however, is simple and is much like the computing technique of the method of determinants, except that the division is by Δ rather than by the preceding pivot. It is illustrated in Table 17.13*a*

TABLE 17.13*a*

CALCULATION OF ADJUGATE FOR ELIMINATED VARIABLES

	a				a_{i5}	x_i
	26	−10	15	32	23	
	19	45	−14	−8	57	
	−12	16	27	13	47	
	32	29	−35	28	−68	
2305327	66233	−16033	42069	−6503	23	2
	56151	28558	33194	−52258	57	1
	−53068	36236	18224	45899	47	3
	−35013	9659	−47056	[53524]	−68	−2
53524	1439	−345	844		23	0.6917
	510	882	−296		57	1.3609
	−535	649	[1360]		47	1.2417
1360	45	−19			23	1.1801
	10	[26]			57	0.7684
26	1				23	0.8846

when the effects of the successive eliminations of $k = 4$, $k = 3$, $k = 2$ are shown in the successive adjugate matrices. The value of a is exhibited to make the presentation complete and for the calculation of Δ, if that is necessary.

The adjugate matrices corresponding to a different order of elimination are shown in Table 17.13*b*, where the order of elimination is $k = 2$, $k = 3$, $k = 4$. The pivotal A'_{kk} terms are indicated.

The elimination of all but one of the variables leads eventually to the result 1, no matter what the order of elimination. This fact can be used as a check of the accuracy of all the adjugates, including the original one. We note, too, that every term results from an exact division so that we have this desirable feature in eliminating mistakes.

The last two columns of Tables 17.13*a* and 17.13*b* give the solutions of the related equations that are obtained by deletion. The first of these columns shows the new value of f and the second shows the values of $x = Af/\Delta$. The numerator is accomplished by a column-by-column multiplication of A' and f. The results of the division can be written

TABLE 17.13b

CALCULATION OF ADJUGATE FOR ELIMINATED VARIABLES—SECOND ILLUSTRATION

	a				f	x_i
$a =$	26	−10	15	32	23	
	19	45	−14	−8	57	
	−12	16	27	13	47	
	32	29	−35	28	−68	
2305327	66233	−16033	42069	−6503	23	2
	56151	[28558]	33194	−52258	57	1
	−53068	36236	18224	45899	47	3
	−35013	9659	−47056	53524	−68	−2
28558	1211		752	−444	23	0.0338
	−1540		[−296]	1390	47	1.8377
	−669		−722	882	−68	−0.1701
−296	28			−32	23	−9.5270
	−32			[26]	−68	−8.459$\bar{5}$
26	1				23	0.8846

in exact fractional form or can be carried to four decimal places as in the illustrations.

The formula (2) may be adjusted to serve as the basis for the corresponding technique that uses the inverse. We use the transpose of the inverse just as we use the adjugate that is the transpose of the adjoint for ease of calculation.

We divide (2) by A' and get

$$\frac{A'_{ij)k(}}{A'_{kk}} = \frac{A'_{ij}}{\Delta} - \frac{A'_{ik}A'_{kj}}{\Delta A'_{kk}} \tag{3}$$

and we have

$$\frac{A'_{ij)k(}}{A'_{kk}} = \frac{A'_{ij}}{\Delta} - \frac{\dfrac{A'_{ik}}{\Delta}\dfrac{A'_{kj}}{\Delta}}{\dfrac{A'_{kk}}{\Delta}}.$$

Now A'_{kk} is the determinant of the matrix, with row k and column k deleted and $c'_{ij} = \dfrac{A'_{ij}}{\Delta}$ so that we have

$$c'_{ij)k(} = c'_{ij} - \frac{c'_{ik}c'_{kj}}{c'_{kk}}. \tag{5}$$

This formula, which is well known [J], can also be obtained directly from (16.2.24). In form it is very similar to the basic formula of the method of single division. The form

$$c'_{ij)k(} = \frac{c'_{kk}c'_{ij} - c'_{ik}c'_{kj}}{c'_{kk}} \tag{6}$$

is recommended for computational use.

Starting with the transpose of the inverse, we use the value of c'_{kk} as a pivot and apply (6). Successive eliminations can be performed. The process is illustrated in Table 17.13*c*, where the calculations for the

TABLE 17.13*c*

CALCULATION OF c' FOR ELIMINATED VARIABLE

a				f	x_i
1.0000	0.4000	0.5000	0.6000	0.2000	
0.4000	1.0000	0.3000	0.4000	0.4000	
0.5000	0.3000	1.0000	0.2000	0.6000	
0.6000	0.4000	0.2000	1.0000	0.8000	
2.0708	−0.1913	−0.7759	−1.0107	0.2000	−0.9365
−0.1913	1.2842	−0.2185	−0.3552	0.4000	0.0602
−0.7759	−0.2185	1.3988	0.2731	0.6000	0.8152
−1.0107	−0.3552	0.2731	[1.6941]	0.8000	1.1749
1.4678	−0.4032	−0.6130		0.2000	−0.2355
−0.4032	1.2097	−0.1612		0.4000	0.3065
−0.6130	−0.1612	[1.3508]		0.6000	0.6234
1.1904	−0.4761			0.2000	0.0476
−0.4761	[1.1905]			0.4000	0.3810
1.0000				0.2000	0.2000

successive eliminations of x_4, x_3, and x_2 are carried. The original equations are stated in the first four rows, and the values of x_i resulting from the eliminations are shown in the last column. Table 17.13*d* is similar

TABLE 17.13*d*

CALCULATION OF c' FOR ELIMINATED VARIABLE

a				f	x_i
1.0000	0.4000	0.5000	0.6000	0.2000	
0.4000	1.0000	0.3000	0.4000	0.4000	
0.5000	0.3000	1.0000	0.2000	0.6000	
0.6000	0.4000	0.2000	1.0000	0.8000	
2.0708	−0.1913	−0.7759	−1.0107	0.2000	−0.9365
−0.1913	[1.2842]	−0.2185	−0.3552	0.4000	0.0602
−0.7759	−0.2185	1.3988	0.2731	0.6000	0.8152
−1.0107	−0.3552	0.2731	1.6941	0.8000	1.1749
2.0423		−0.8084	−1.0636	0.2000	−0.9275
−0.8084		[1.3616]	0.2127	0.6000	0.8254
−1.0636		0.2127	1.5959	0.8000	1.0321
1.5623			−0.9373	0.2000	−0.4374
−0.9373			[1.5627]	0.8000	1.0627
1.0001				0.2000	0.2000

to Table 17.13*c*, except that the elimination is in x_2, x_3, x_4 order. In both cases the elimination process, as applied to the transpose of the inverse, leads approximately to unity. This fact may be used as a check of the approximate correctness of the original inverse and those obtained from it by deletion.

17.14 Additional references. Additional material of general interest may be found in references [K].

REFERENCES

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

B. W. E. Milne, *Numerical Calculus*, Princeton University Press, 1949, pp. 29–35.

C. F. E. Satterthwaite, "Error control in matrix calculation," *Annals of Mathematical Statistics*, **15**, 373–387 (1944).

D. I. M. H. Etherington, "On errors in determinants," *Proceedings of the Edinburgh Mathematical Society*, Series 2, **3**, 107–117 (1932).

E. R. S. Burington, *Handbook of Mathematical Tables and Formulas*, Handbook Publishers, Inc., Sandusky, Ohio, Third Edition, 1948. See p. 5.

F. F. A. Willers (translated by R. T. Beyer), *Practical Analysis*, Dover Publications, New York, pp. 271–273.

G. 1. P. S. Dwyer and M. S. Macphail, "Symbolic matrix derivatives," *Annals of Mathematical Statistics*, **19**, 517–534 (1948). See section 12.
2. A. D. Michal, *Matrix and Tensor Calculus*, John Wiley and Sons, New York, 1947.

H. 1. H. Hotelling, "Some new methods in matrix calculation," *Annals of Mathematical Statistics*, **14**, 1–33 (1943).
2. A. M. Turing, "Rounding off errors in matrix processes," *Quarterly Journal of Mechanics and Applied Mathematics*, **1**, 287–308 (1948).

I. J. Sherman and Winifred Morrison, "Adjustment of an inverse matrix corresponding to a change in one element of a given matrix," Abstract, *Annals of Mathematical Statistics*, **20**, 317 (1949).

J. 1. R. A. Fisher, *Statistical Methods for Research Workers*, Tenth Edition, Oliver and Boyd, London and Glasgow, 1944. Section 29.1.
2. W. G. Cochran, "The omission or addition of an independent variate in multiple linear regression," *Supplement Journal Royal Statistical Society*, **5**, 171–176 (1938).

Additional material of interest may be found in:

K. 1. F. R. Moulton, "On the solution of equations having small determinants," *American Mathematical Monthly*, **20**, 242–249 (1913).
2. L. B. Tuckerman, "On the mathematically significant figures in the solution of simultaneous linear equations," *Annals of Mathematical Statistics*, **12**, 307–316 (1941).
3. A. T. Lonseth, "Systems of linear equations with coefficients subject to error," *Annals of Mathematical Statistics*, **13**, 332–337 (1942).
4. A. W. Wundheiler, "The necessity of error analysis in numerical computation," *The Annals of the Computation Laboratory of Harvard University*, **16**, 83–90.
5. H. A. Rademacher, "On the accumulation of errors in processes of integration on high speed calculating machines," *The Annals of the Computation Laboratory of Harvard University*, **16**, 176–187 (1948).
6. D. H. Leavens, "Accuracy in the Doolittle solution," *Econometrika*, **15**, 45–50 (1947).

EXERCISES

1–5. Find the maximum first-order errors for the determinants of exercises 1 to 5 of Chapter 9 if $|\epsilon_{ij}| \leq 0.005$. Use the method of determinants.

6–8. Find the maximum first-order errors for the determinants of exercises 9.4, 9.5, and 9.11 by the compact method of single division.

9. Use the method of Table 17.7*a* in finding the maximum errors of the solution of exercise 4.1 if $\eta = 0.005$.

10. Use the method of Table 17.7*a* in finding the maximum errors of the solution of exercise 4.10 if $\eta = 0.005$.

11. Use the method of Table 17.7*b* in finding the maximum errors of the solution of exercise 5.4 if $\eta = 0.0005$.

12. Find the maximum errors in the solution of exercise 4.1 for x, y, and z if $\eta = 0.01$. Derive the sets of equations having these maximum errors.

13. Use the method described by Milne in getting bounds for the errors of the solution of the problem of exercise 4.1 if $\eta = 0.01$.

14. Use the method described by Willers in getting bounds for the errors of the solution of the problem of exercise 4.1 if $\eta = 0.01$.

15. Use the method of Table 17.8*a* in finding the maximum possible errors of the solution of 4.1 if $|\epsilon(f)| \leq 0.005$ and $\epsilon(a) = 0$.

16. Work exercise 15 by the method of Table 17.8*b*.

17. Work exercise 15 with $\epsilon(f) = 0$ and $|\epsilon(a)| \leq 0.005$ by the method of Table 17.8*c*.

18. Work exercise 17 by the method of Table 17.8*d*.

19. Suppose the coefficients of y in exercise 4.1 are subject to errors that are not greater in absolute value than 0.005. The other coefficients are not subject to error. Use the methods of Table 17.8*e* and the method of Table 17.8*f* in finding maximum errors in the result.

20. Work exercise 19 where the errors are not those of the coefficients of y but are those of the coefficients of all the terms in the second row, including the error of the right side of the second equation.

21. Determine maximum errors for the elements of the adjugate of the matrix of coefficients of Table 4.2*a* if $\eta = 0.005$.

22. Determine maximum errors for the elements of the inverse of the matrix of coefficients of Table 4.2*a* if $\eta = 0.005$.

23. Use the method of Table 17.11*a* in solving the equations of exercise 6.11.

24. Make a statement of the problem of Table 17.12*a* and work out the details of the solution.

25. Use the method of Table 17.13*b* in eliminating variables from the adjoint of the matrix of the coefficients of exercise 4.10.

26. Use the method of Table 17.13*d* in eliminating variables from the inverse of the matrix of the coefficients of exercise 6.11.

27. Show $(a - s)^{-1} - a^{-1} = [(I - a^{-1}s)^{-1} - I]a^{-1}$. The first-order approximation (in s) is then $a^{-1}sa^{-1}$. This formula is given by Turing [H.2]. Show that it is the equivalent of (17.9.2). Show also (Turing) that a bound for any term in the inverse is $n^2M(a^{-1})^2\epsilon$ where a is an n by n matrix, $M(a^{-1})$ is the element of a^{-1} having the largest absolute value, and ϵ is a bound for the error of all elements of a.

CHAPTER 18

Application to Statistics

18.1 Introduction. In this chapter specific applications are made to statistics in order to familiarize the reader with their use in solving representative problems. No exhaustive treatment is attempted. The reader who understands the theory of earlier chapters and who has a knowledge of the basic statistical concepts should need only a few illustrations to point the way toward improved calculational procedure. The emphasis here is on the improvement of the computational design and not on the statistical problem itself, nor on its interpretation.

Problems in the field of statistics have, of course, been mentioned in earlier chapters. For instance, the calculation of the sum of squares, the calculation of the correlation coefficient, and the computation of a_3 are indicated in Chapter 3. In addition, many of the solutions of the later chapters have been designed to deal with least squares problems.

18.2 The large variance. The formulas of Chapter 3 feature such expressions as $N\Sigma X^2 - (\Sigma X)^2$ and $N\Sigma XY - (\Sigma X)(\Sigma Y)$. Since these are N times the variance and covariance, respectively, it seems appropriate to call these the *large variance* and the *large covariance.* If we indicate the first by L_{xx} and the second by L_{xy}, (3.2.2) and (3.2.4) become

$$\Sigma x^2 = \frac{L_{xx}}{N} \tag{1}$$

$$\rho_{xy} = \frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}}. \tag{2}$$

In case a single variable is under consideration we may drop the subscripts and indicate the large variance by the symbol L. Thus (1) could appear as

$$\Sigma x^2 = \frac{L}{N}. \tag{3}$$

The chief reason for introducing the large variance is its property of exactness. It is especially useful in work involving the analysis of

variance since the results can be checked exactly and many of the formulas are simpler than those involving the sum of squares of deviates.

18.3 The avoidance of approximations. The use of the large variance enables us to avoid the many divisions of the conventional methods. Also we can eliminate, or at least postpone, many divisions in the formulas used in tests of significance. This is usually accomplished by multiplying the numerator and the denominator by the least common denominator of the divisors. Consider the formula for Student-Fisher t,

$$t = \frac{\bar{X} - \mu}{s}\sqrt{N - 1}, \tag{1}$$

where N is the sample size, $\bar{X}$ is the sample mean, s is the historical s used by Student and is the square root of the second sample moment about its mean, and μ is the population mean, known by specification or hypothesis. Algebraic manipulation results in

$$t = \frac{(\Sigma X - N\mu)}{\sqrt{L}}\sqrt{N - 1}, \tag{2}$$

where ΣX is the sample sum.

When $\mu = 0$, we have

$$t = \frac{\Sigma X}{\sqrt{L}}\sqrt{N - 1}. \tag{3}$$

If $N = 10$, $\Sigma X = 15.8$, $\Sigma X^2 = 38.58$, as in the illustration used by Fisher [A], then $L_{xx} = 1.3616$ and $t = 4.06$.

It is also possible to avoid the approximation caused by taking a square root in (2) and (3) by squaring each side and obtaining

$$t^2 = \frac{(\Sigma X - N\mu)^2}{L}(N - 1) \tag{4}$$

and, if $\mu = 0$,

$$t^2 = \frac{(\Sigma X)^2(N - 1)}{L}. \tag{5}$$

This value can be interpreted by using an F table, with $n_1 = 1$ and $n_2 = N - 1$ degrees of freedom. In the illustration above

$$t^2 = \frac{(15.8)^2 9}{1.3616} = 16.50$$

and this is significant at the 1% level.

The problem and its solution appear in Table 18.3*a*.

TABLE 18.3*a*

CALCULATION OF t^2 AND t

General	Illustration
X_1	1.2
X_2	2.4
X_3	1.3
X_4	1.3
X_5	0.0
X_6	1.0
X_7	1.8
X_8	0.8
X_9	4.6
X_{10}	1.4
N	10
ΣX	15.8
ΣX^2	38.58
L	1.3616
t^2	16.50
t	4.06

The usual formula for testing whether two samples could come from the same normal population becomes, when subjected to appropriate algebraic treatment,

$$t = \frac{N_2\Sigma X_1 - N_1\Sigma X_2}{\sqrt{N_2L_1 + N_1L_2}} \sqrt{\frac{N_1 + N_2 - 2}{N_1 + N_2}} \tag{6}$$

or

$$t^2 = \frac{(N_2\Sigma X_1 - N_1\Sigma X_2)^2}{(N_2L_1 + N_1L_2)} \frac{(N_1 + N_2 - 2)}{(N_1 + N_2)}. \tag{7}$$

These values are easily computed from a computing form that features N_i, ΣX_i, and L_i. Application is made in Table 18.3*b* to the problem [A], which has been used by Fisher and others.

If $N_1 = N_2$ the formula (7) becomes

$$t^2 = \frac{(\Sigma X_1 - \Sigma X_2)^2(N - 1)}{L_1 + L_2}. \tag{8}$$

TABLE 18.3b

COMPUTATION OF t^2 AND t FOR DIFFERENCE OF TWO MEANS

	x_1	x_2
	0.7	1.9
	−1.6	0.8
	−0.2	1.1
	−1.2	0.1
	−0.1	−0.1
	3.4	4.4
	3.7	5.5
	0.8	1.6
	0.0	4.6
	2.0	3.4
N_i	10	10
ΣX_1	7.5	23.3
ΣX_i^2	34.43	90.37
L_i	288.05	360.81
$N_1\Sigma X_2 - N_2\Sigma X_1$	158	
$N_1L_2 + N_2L_1$	6488.6	
Numerator of t^2	449352	
Denominator of t^2	129772	
t^2	3.463	
t	1.861	

18.4 Regression tests. In this section application of improved computing methods is made to some of the usual tests for regression. Thus the usual test for significance of a regression is

$$t = \frac{b - \beta}{\sqrt{\frac{\Sigma E^2}{N-2}\frac{1}{\Sigma x^2}}} = (b - \beta)\sqrt{\frac{L_{xx}}{L_{EE}}}\sqrt{N-2}, \tag{1}$$

where

$$b = \frac{L_{xy}}{L_{xx}}.$$

Now it can be shown that

$$L_{EE} = L_{yy} - bL_{xy} = \frac{L_{xx}L_{yy} - L_{xy}^2}{L_{xx}} = \frac{\Delta}{L_{xx}},$$

where

$$\Delta = \begin{vmatrix} L_{xx} & L_{xy} \\ L_{xy} & L_{yy} \end{vmatrix}.$$

It follows that

$$t = (b - \beta)L_{xx}\sqrt{\frac{N-2}{\Delta}} \tag{2}$$

and hence that

$$t^2 = \frac{(b-\beta)L_{xx}^2(N-2)}{L_{xx}L_{yy} - L_{xy}^2}. \tag{3}$$

The hypothesis that $\beta = 0$ is commonly tested. In this case we have, since $b = \dfrac{L_{xy}}{L_{xx}}$,

$$t^2 = \frac{L_{xy}^2(N-2)}{L_{xx}L_{yy} - L_{xy}^2}. \tag{4}$$

Application is made in Table 18.4*a* to a trivial problem proposed by Rider [B]. The general notation is presented on the left and the illus-

TABLE 18.4*a*

TEST FOR REGRESSION COEFFICIENT

General			Illustration		
i	X_i	Y_i	i	X_i	Y_i
1	X_1	Y_1	1	0	1
2	X_2	Y_2	2	1	3
. . . . .			3	3	2
			4	6	5
N	X_N	Y_N	5	8	4
N	ΣX	ΣY	5	18	15
	ΣX^2	ΣXY		110	71
		ΣY^2			55
	L_{xx}	L_{xy}		226	85
		L_{yy}			50
a	b	Δ	1.646	0.376	4075
		t^2			5.319
		t			2.306

tration on the right. The use of a table of F or t shows the coefficient significant at the 5% level. The test for the difference between regression coefficients is the Student-Fisher t.

$$(5) \qquad t = \frac{b_1 - b_2}{\sqrt{\left(\frac{1}{\Sigma {x_1}^2} + \frac{1}{\Sigma {x_2}^2}\right)\left(\frac{\Sigma {E_1}^2 + \Sigma {E_2}^2}{N_1 + N_2 - 4}\right)}}$$

$$= \frac{b_1 - b_2}{\sqrt{\frac{N_1}{L_{xx\cdot 1}} + \frac{N_2}{L_{xx\cdot 2}}}} \sqrt{\frac{N_1 + N_2 - 4}{\frac{\Delta_1}{N_1 L_{xx\cdot 1}} + \frac{\Delta_2}{N_2 L_{xx\cdot 2}}}},$$

where

$$\Delta_1 = \begin{vmatrix} L_{xx\cdot 1} & L_{xy\cdot 1} \\ L_{xy\cdot 1} & L_{yy\cdot 1} \end{vmatrix} \quad \text{and} \quad \Delta_2 = \begin{vmatrix} L_{xx\cdot 2} & L_{xy\cdot 2} \\ L_{xy\cdot 2} & L_{yy\cdot 2} \end{vmatrix}.$$

The values of Δ_1/N_1 and Δ_2/N_2 are exact and can be recorded as Δ'_1 and Δ'_2. We then have

$$(6) \qquad {t^2}_{b_1 - b_2} = \frac{(b_1 - b_2)^2 (L_{xx\cdot 1} L_{xx\cdot 2})^2 (N_1 + N_2 - 4)}{(N_1 L_{xx\cdot 2} + N_2 L_{xx\cdot 1})(L_{xx\cdot 1}\Delta'_2 + L_{xx\cdot 2}\Delta'_1)},$$

TABLE 18.4*b*

TEST FOR DIFFERENCE OF REGRESSION COEFFICIENTS—COMPUTING FORM

$i \quad X_1 \quad Y_1$ . . . . . . $N_1 \quad X_{1N_1} \quad Y_{1N_1}$	$i \quad X_2 \quad Y_2$ . . . . . . $N_2 \quad X_{2N_2} \quad Y_{2N_2}$
$N_1 \quad \Sigma X_1 \quad \Sigma Y_1$ $\Sigma {X_1}^2 \quad \Sigma X_1 Y_1$ $\Sigma {Y_1}^2$	$N_2 \quad \Sigma X_2 \quad \Sigma Y_2$ $\Sigma {X_2}^2 \quad \Sigma X_2 Y_2$ $\Sigma {Y_2}^2$
$L_{xx\cdot 1} \quad L_{xy\cdot 1}$ $L_{yy\cdot 1}$	$L_{xx\cdot 2} \quad L_{xy\cdot 2}$ $L_{yy\cdot 2}$
$b_1 \quad \Delta'_1$	$b_2 \quad \Delta'_2$
	$L_{xx\cdot 2} L_{xy\cdot 1} - L_{xx\cdot 1} L_{xy\cdot 2}$ $N_1 + N_2 - 4 = DF$ $L_{xx\cdot 1}\Delta'_2 + L_{xx\cdot 2}\Delta'_1$ $L_{xx\cdot 2} N_1 + L_{xx\cdot 1} N_2$ Numerator of t^2 Denominator of t^2 t^2

where all the values except b_1 and b_2 are exact. Substitution of the values of b_1 and b_2 gives

$$t^2_{b_1-b_2} = \frac{(L_{xx\cdot 2}L_{xy\cdot 1} - L_{xx\cdot 1}L_{xy\cdot 2})^2(N_1 + N_2 - 4)}{(N_1L_{xx\cdot 2} + N_2L_{xx\cdot 1})(L_{xx\cdot 1}\Delta'_2 + L_{xx\cdot 1}\Delta'_1)}, \tag{7}$$

where all the values are exact. See Tables 18.4*b* and 18.4*c*.

TABLE 18.4*c*

TEST FOR DIFFERENCE OF REGRESSION COEFFICIENTS

i_1	X_1	Y_1	i_2	X_2	Y_2
1	0	1	1	0	1
2	1	3	2	1	2
3	3	2	3	2	2
4	6	5	4	3	2
5	8	4	5	4	3
			6	5	4
			7	6	5
			8	7	6
5	18	15	8	28	25
	110	71		140	116
		55			99
	226	85		336	228
		50			167
	0.376	815		0.679	516
				−22968	
				9	
				390456	
				3488	
				$4748 \cdot 10^6$	
				$1362 \cdot 10^6$	
				3.49	

not proved significant

The formula (7) simplifies slightly when $N_1 = N_2$. If in addition the values of Y_1 and Y_2 are related to the same N values of X, $L_{xx\cdot 1} = L_{xx\cdot 2}$ and we have

$$t^2_{b_1-b_2} = \frac{(L_{xy\cdot 1} - L_{xy\cdot 2})^2(N-2)}{\Delta'_1 + \Delta'_2} \tag{8}$$

which has considerable similarity to (18.3.8).

It is not proper here to take space to present appropriate formulas for a variety of regression problems.

18.5 Analysis of variance. Exact methods can also be used in many problems involving the analysis of variance. Application is first made to a two-way classification problem without replications. The basic identity becomes an addition formula in the large variance rather than in the sums of squared deviates. Thus if i indicates the row and j the column, the fundamental identity becomes

$$L_{ij} = L_{i\cdot} + L_{\cdot j} + I_{ij} \tag{1}$$

where L_{ij} is the large variance of the individual elements, $L_{i\cdot}$ is the large variance of the row sums, $L_{\cdot j}$ is the large variance of the column sums, and I_{ij} is the residual large variance. We associate with each of these terms the usual degrees of freedom and are able to make the F tests, using the formula

$$F = \frac{\text{large variance}}{DF} \div \frac{\text{large variance}}{DF}. \tag{2}$$

A computing form for the terms of (1) is shown in Table 18.5a.

TABLE 18.5a

COMPUTING FORM FOR ANALYSIS OF VARIANCE

TWO-WAY CLASSIFICATION WITHOUT REPLICATIONS

X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	$X_{1\cdot}$		
X_{21}	X_{22}	X_{23}	X_{24}	X_{25}	$X_{2\cdot}$		
X_{31}	X_{32}	X_{33}	X_{34}	X_{35}	$X_{3\cdot}$		
X_{41}	X_{42}	X_{43}	X_{44}	X_{45}	$X_{4\cdot}$		
$X_{\cdot 1}$	$X_{\cdot 2}$	$X_{\cdot 3}$	$X_{\cdot 4}$	$X_{\cdot 5}$	$X_{\cdot\cdot}$	$\Sigma X_{\cdot j}^2$	$L_{\cdot j}$
					$\Sigma X_{i\cdot}^2$	ΣX_{ij}^2	I_{ij}
					$L_{i\cdot}$	I_{ij}	L_{ij}

The values of $X_{i\cdot}$ are row sums and those of $X_{\cdot j}$ are column sums. The value of $X_{\cdot\cdot}$ is the total of the individual items. If there are a

rows and b columns, we have

$$
\begin{aligned}
L_{ij} &= ab\Sigma X_{ij}^2 - X_{..}^2 \\
L_{i\cdot} &= a\Sigma X_{i\cdot}^2 - X_{..}^2 \\
L_{\cdot j} &= b\Sigma X_{\cdot j}^2 - X_{..}^2 \\
I_{ij} &= L_{ij} - L_{i\cdot} - L_{\cdot j}.
\end{aligned} \tag{3}
$$

The formulas (3) alone are needed. An additional computation of the value I_{ij} may be obtained from the formula

$$I_{ij} = ab\Sigma X_{ij}^2 - a\Sigma X_{i\cdot}^2 - b\Sigma X_{\cdot j}^2 + X_{..}^2. \tag{4}$$

The computing form is applied to a problem taken from Snedecor's Table 11.1 [C] in Table 18.5*b*.

TABLE 18.5*b*

HEIGHTS (CENTIMETERS) OF SOY PLANTS IN FIVE WEEKS

FOUR BLOCKS OF REPLICATIONS

Week								
Block	1	2	3	4	5			
	4	18	26	38	44	130		
	3	19	25	35	43	125		
	6	18	24	28	39	115		
	7	13	21	31	38	110		
	20	68	96	132	164	480	58560	62400
						57850	14770	1600
						1000	1600	65000

A further analysis similar to the conventional type is made in Table 18.5*c*, where the values of N, ΣX, ΣX^2, L, L or I, DF, and the usual sum of squares are indicated for the various sources. The use of an L or an I column in addition to the L column permits us to list the L_{ij} terms and the I_{ij} terms in the same row.

Table 18.5*c* is an explanatory table and should be replaced by the form of Table 18.5*d* for actual computation of F. The first four columns show just how the L's of Table 18.5*b* are calculated. The last column shows that the conventional sum of squares can be obtained by dividing

the L or I column by the total number of observations. This is not necessary, however, since (2) can be used as the basis of the calculation of F. The calculation of F is shown in Table 18.5*d*. The values

TABLE 18.5*c*

RELATION OF TERMS IN ANALYSIS OF VARIANCE

Source	N	ΣX	ΣX^2	L	L or I	DF	Sum of Squares
Weeks	5	480	58560	62400	62400	4	3120
Blocks	4	480	57850	1000	1000	3	50
Weeks—blocks	20	480	14770	65000	1600	12	80

are inserted from Table 18.5*b*, and the values of F are computed directly.

We may use a computing form that is more general than that of Table 18.5*a*. Such a form is shown in Table 18.5*e*, where the large

TABLE 18.5*d*

CALCULATION OF F

Source	L or I	DF	$\frac{L \text{ or } I}{DF}$	F
Weeks	62400	4	15600	117
Blocks	1000	3	333	2.5
Weeks—blocks	1600	12	133	1

TABLE 18.5*e*

COMPUTING FORM FOR ANALYSIS OF VARIANCE WITH CHOICES

TWO-WAY CLASSIFICATION WITHOUT REPLICATIONS

X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	$X_{1\cdot}$	ΣX_{i1}^2	$L_{1\cdot}$
X_{21}	X_{22}	X_{23}	X_{24}	X_{25}	$X_{2\cdot}$	ΣX_{i2}^2	$L_{2\cdot}$
X_{31}	X_{32}	X_{33}	X_{34}	X_{35}	$X_{3\cdot}$	ΣX_{i3}^2	$L_{3\cdot}$
X_{41}	X_{42}	X_{43}	X_{44}	X_{45}	$X_{4\cdot}$	ΣX_{i4}^2	$L_{4\cdot}$
$X_{\cdot 1}$	$X_{\cdot 2}$	$X_{\cdot 3}$	$X_{\cdot 4}$	$X_{\cdot 5}$	$X_{\cdot\cdot}$	$\Sigma X_{\cdot j}^2$	$L_{\cdot j}$
ΣX_{i1}^2	ΣX_{i2}^2	ΣX_{i3}^2	ΣX_{i4}^2	ΣX_{i5}^2	$\Sigma X_{i\cdot}^2$	$\Sigma X_{\cdot j}^2$	I_{ij}
$L_{\cdot 1}$	$L_{\cdot 2}$	$L_{\cdot 3}$	$L_{\cdot 4}$	$L_{\cdot 5}$	$L_{i\cdot}$	I_{ij}	L_{ij}

TABLE 18.5*g*

ILLUSTRATION OF TABLE 18.5*f*—GOULDEN

9.4	2.6	12.3	4.6	13.5	42.4
9.6	3.1	13.0	4.3	13.8	43.8
9.6	2.7	12.4	1.8	13.0	39.5
28.6	8.4	37.7	10.7	40.3	125.7

13.7	21.6	19.4	13.5	24.5	92.7
12.7	22.6	20.6	10.4	24.3	90.6
12.6	21.8	20.9	6.8	23.2	85.3
39.0	66.0	60.9	30.7	72.0	268.6

23.1	24.2	31.7	18.1	38.0	135.1
22.3	25.7	33.6	14.7	38.1	134.4
22.2	24.5	33.3	8.6	36.2	124.8
67.6	74.4	98.6	41.4	112.3	394.3

Source	N	ΣX	ΣX^2	L	L or I	DF
Row	3	394.3	51890.41	198.74	198.74	2
Column	5	394.3	34152.33	15289.16	15289.16	4
Layer	2	394.3	87946.45	20420.41	20420.41	1
$R \times C$	15	394.3	11436.57	16076.06	588.16	8
$R \times L$	6	394.3	29354.19	20652.65	33.50	2
$C \times L$	10	394.3	19760.69	42134.41	6424.84	4
$R \times C \times L$	30	394.3	6618.43	43080.41	125.60	8

TABLE 18.5*f*

COMPUTING FORM FOR A THREE-WAY CLASSIFICATION

X_{111}	X_{121}	X_{131}	X_{141}	X_{151}	$X_{1\cdot 1}$
X_{211}	X_{221}	X_{231}	X_{241}	X_{251}	$X_{2\cdot 1}$
X_{311}	X_{321}	X_{331}	X_{341}	X_{351}	$X_{3\cdot 1}$
$X_{\cdot 11}$	$X_{\cdot 21}$	$X_{\cdot 31}$	$X_{\cdot 41}$	$X_{\cdot 51}$	$X_{\cdot\cdot 1}$

X_{112}	X_{122}	X_{132}	X_{142}	X_{152}	$X_{1\cdot 2}$
X_{212}	X_{222}	X_{232}	X_{242}	X_{252}	$X_{2\cdot 2}$
X_{312}	X_{322}	X_{332}	X_{342}	X_{352}	$X_{3\cdot 2}$
$X_{\cdot 12}$	$X_{\cdot 22}$	$X_{\cdot 32}$	$X_{\cdot 42}$	$X_{\cdot 52}$	$X_{\cdot\cdot 2}$

$X_{11\cdot}$	$X_{12\cdot}$	$X_{13\cdot}$	$X_{14\cdot}$	$X_{15\cdot}$	$X_{1\cdot\cdot}$
$X_{21\cdot}$	$X_{22\cdot}$	$X_{23\cdot}$	$X_{24\cdot}$	$X_{25\cdot}$	$X_{2\cdot\cdot}$
$X_{31\cdot}$	$X_{32\cdot}$	$X_{33\cdot}$	$X_{34\cdot}$	$X_{35\cdot}$	$X_{3\cdot\cdot}$
$X_{\cdot 1\cdot}$	$X_{\cdot 2\cdot}$	$X_{\cdot 3\cdot}$	$X_{\cdot 4\cdot}$	$X_{\cdot 5\cdot}$	$X_{\cdot\cdot\cdot}$

Source	N	ΣX	ΣX^2	L	L or I	DF
i	3					2
j	5					4
h	2					1
if	15					8
ih	6					2
jh	10					4
ijh	30					8

TABLE 18.5*g*

ILLUSTRATION OF TABLE 18.5*f*—GOULDEN

9.4	2.6	12.3	4.6	13.5	42.4
9.6	3.1	13.0	4.3	13.8	43.8
9.6	2.7	12.4	1.8	13.0	39.5
28.6	8.4	37.7	10.7	40.3	125.7

13.7	21.6	19.4	13.5	24.5	92.7
12.7	22.6	20.6	10.4	24.3	90.6
12.6	21.8	20.9	6.8	23.2	85.3
39.0	66.0	60.9	30.7	72.0	268.6

23.1	24.2	31.7	18.1	38.0	135.1
22.3	25.7	33.6	14.7	38.1	134.4
22.2	24.5	33.3	8.6	36.2	124.8
67.6	74.4	98.6	41.4	112.3	394.3

Source	N	ΣX	ΣX^2	L	L or I	DF
Row	3	394.3	51890.41	198.74	198.74	2
Column	5	394.3	34152.33	15289.16	15289.16	4
Layer	2	394.3	87946.45	20420.41	20420.41	1
$R \times C$	15	394.3	11436.57	16076.06	588.16	8
$R \times L$	6	394.3	29354.19	20652.65	33.50	2
$C \times L$	10	394.3	19760.69	42134.41	6424.84	4
$R \times C \times L$	30	394.3	6618.43	43080.41	125.60	8

variances are computed for each column and each row. Additional checking formulas are

$$
\begin{aligned}
L_{ij} &= a[L_{\cdot 1} + L_{\cdot 2} + \cdots + L_{\cdot a}] + L_{\cdot j} \\
L_{ij} &= b[L_{1\cdot} + L_{2\cdot} + \cdots + L_{b\cdot}] + L_{i} \\
I_{ij} &= a[L_{\cdot 1} + L_{\cdot 2} + \cdots + L_{\cdot a}] - L_{i\cdot} \\
I_{ij} &= b[L_{1\cdot} + L_{2\cdot} + \cdots + L_{b\cdot}] - L_{\cdot j}.
\end{aligned} \tag{5}
$$

The large variance can be used in problems involving more classifications when the number of replications is constant. The basic large variance formula is nothing but the conventional identity in the sum of squares multiplied by the number of observations. In a three-way classification without replications it is

$$L_{ijh} = L_{i\cdot\cdot} + L_{\cdot j\cdot} + L_{\cdot\cdot h} + I_{ij\cdot} + I_{i\cdot h} + I_{\cdot jh} + I_{ijh}.$$

Application is made to a three-way classification problem used as an illustration by Goulden [D]. The general form of the presentation is given in Table 18.5*f* and the numerical illustration in Table 18.5*g*.

The exact values of the sum of squares approximated by Goulden can be computed exactly by dividing the L or I column of Table 18.5*f* by 30. This general technique is applicable to a variety of problems in the analysis of variance.

18.6 Deviates. It is conventional to use the deviation from the mean, the *deviate*

$$x_i = X_i - \overline{X}, \tag{1}$$

or the standard deviate

$$t_i = \frac{X_i - \overline{X}}{\sigma_x}, \tag{2}$$

as the basis of theoretical work in certain branches of statistics. Each of these is good for theoretical purposes, but neither is very satisfactory for practical purposes since the direct calculation demands the approximate operation of division. It is customary to avoid this division in part by indirect methods, but even these usually feature the (approximate) mean.

A more satisfactory method, from the standpoint of approximation, results when (1) is put in the form

$$x = \frac{NX_i - \Sigma X}{N} \tag{3}$$

and the division is postponed until the final steps, as in (3.4.2).

Now the numerator of (3), which is exact for digital values of X_i, may be used as the basis of the calculation with a final adjustment using the formulas for multiplication by a constant. Thus if we call $NX_i - \Sigma X$ the *large deviate*, D_i, we can compute the sum of the squares of D_i and divide by N to get the conventional sum of the squares. Thus

$$\Sigma D_i^2 = \Sigma(NX_i - \Sigma X)^2 = N[N\Sigma X_i^2 - (\Sigma X_i)^2] \tag{4}$$

so that

$$\Sigma(X_i - \overline{X})^2 = \frac{N\Sigma X^2 - (\Sigma X)^2}{N}$$

as in (3.4.2).

Since the standard moments are independent of a change in scale, it follows that α_3 can be computed directly from the powers of the large deviate. This is substantially what the technique for the calculation of α_3 in Chapter 3 is.

The same holds true for correlation problems. Formula (3.3.4), for example, may be viewed as the correlation between the large deviates $NX_i - \Sigma X$ and $NY_i - \Sigma Y$.

The technique suggested for the analysis of variance in the last section may also be viewed similarly as the analysis of the sum of the squares of the large deviate. The resulting formulas are much simpler, as each of the possible terms has a coefficient of unity and the formulas are exact.

It is sometimes feasible, for theoretical purposes, to divide the standard deviate by $\sqrt{N}$ to get a deviate that has the property

$$\frac{\Sigma(X - \overline{X})^2}{\sqrt{N}\sigma_x} = 1$$

and

$$\frac{\Sigma(X - \overline{X})}{\sqrt{N}\sigma_x}\frac{(Y - \overline{Y})}{\sqrt{N}\sigma_y} = \rho.$$

This might be called the *small standard deviate.*

In addition to the deviate and the standard deviate, we might have the large and small deviate and the large and small standard deviate. The definition of each of these together with the values of the sum of the squares of the deviates and the products of the deviates of X and Y are given in Table 18.6*a*.

In deriving theory we may let x be any of the six deviates defined in Table 18.6*a*. Interpretation of the results can then be made according to the definition. This is frequently done with the deviate and the standard deviate and it can be extended to take care of the other types

of deviates that we might wish to use, either for theoretical or for computational purposes.

Thus the results of a matrix presentation of least squares theory, such as that of [E], can then be expressed in the language of L's with the use

TABLE 18.6a

FORMULAS FOR DEVIATES

Type of Deviate	Definition	Sum of Squares	Sum of Products
Large deviate	$NX - \Sigma X$	$NL_{xx} = N^3\sigma_x^2$	$NL_{xy} = N^3\sigma_x\sigma_y\rho_{xy}$
Deviate	$X - \bar{X}$	$\Sigma(X - \bar{X})^2 = N\sigma_x^2$	$N\sigma_x\sigma_y\rho_{xy}$
Small deviate	$\frac{X - \bar{X}}{\sqrt{N}}$	σ_x^2	$\sigma_x\sigma_y\rho_{xy}$
Large standard deviate	$\frac{NX - \Sigma X}{\sigma_x}$	N^3	$N^3\rho_{xy}$
Standard deviate	$\frac{X - \bar{X}}{\sigma_x}$	N	$N\rho_{xy}$
Small standard deviate	$\frac{X - \bar{X}}{\sqrt{N}\sigma_x}$	1	ρ_{xy}

of the large deviate, or in the language of correlation coefficients with the use of the small standard deviate.

18.7 Application to least squares. A whole book could be written on the application of the theory of the earlier chapters to least squares. Detailed applications are not presented here as the mathematical material developed closely parallels that of least squares, but the theory of least squares in matrix form [E] (or in some equivalent form [F]) follows readily. Duncan and Kenney [G] have prepared a monograph that features the square root method.

18.8 Multiple correlation and regression problems. Extensive treatment could be given to multiple correlation and regression problems. The author has written a number of papers in which the techniques are the direct application of the methods of this treatise. The reader who is familiar with the conventional theory of multiple and partial correlation and regression and who understands the methods described in the

earlier chapters should have little trouble in applying them to the more specific material. We should bear in mind that the use of alternative deviates makes possible the presentation of the calculation in a variety of forms. Thus it is possible to study the general subject of multiple and partial correlation and regression with the use of the large deviate without the necessity of making the customary transformations that result in formulas that are the equivalents of those obtained with the use of the small standard deviate. Many of these formulas have been developed by the author, who has not yet published them. The general theory, which is the basis of these formulas, is explained in earlier chapters of this book, and the understanding reader should be able to draw up his own improved calculational techniques. The method of determinants should play a very important role in this since with it conventional determinantal results can be translated into suitable form. The type of matrix manipulation featured in Chapters 12 to 15 is also of assistance.

Before leaving this topic we note that the conventional tests for significance of a standard regression coefficient involve the diagonal elements of the inverse. These are easily computed when the equations are solved with the reduction of I. An illustration used earlier by the author [H] is presented in Table 18.8*a*, where the standard normal

TABLE 18.8*a*

SUCCESSIVE CORRELATION AND REGRESSION CONSTANTS

1.000	0.313	0.280	0.182	0.166	0.495	1	0	0	0	0	x_6 = weight
*	1.000	0.552	0.554	0.615	0.650	0	1	0	0	0	N = 1000
*	*	1.000	0.747	0.693	0.803	0	0	1	0	0	$\bar{x}_6$ = 139.3 lb.
*	*	*	1.000	0.774	0.804	0	0	0	1	0	σ_6 = 17.1 lb.
*	*	*	*	1.000	0.312	0	0	0	0	1	
1.000	0.313	0.280	0.182	0.166	0.495	1.000	0	0	0	0	x_1 = height
	0.950	0.594	0.523	0.593	0.521	−0.329	1.053	0	0	0	x_2 = shoulder girth
		0.754	0.511	0.390	0.471	−0.112	−0.830	1.326	0	0	x_3 = chest girth
			0.657	0.357	0.306	0.072	−0.193	−1.031	1.522	0	x_4 = waist girth
				0.584	0.219	0.081	−0.397	−0.255	−0.930	1.712	x_5 = right thigh girth
0.495					0.755	1.000					0.869 14.9 lb. 0.495
0.324	0.549				0.484	1.108	1.109				0.696 11.9 lb. 0.718
0.271	0.158	0.625			0.262	1.121	1.798	1.753			0.512 8.8 lb. 0.859
0.293	0.099	0.309	0.466		0.168	1.126	1.835	2.821	2.316		0.410 7.0 lb. 0.912
0.311	0.012	0.253	0.262	0.375	0.120	1.133	1.993	2.886	3.181	2.931	0.346 5.9 lb. 0.938

equations are solved with the square root method. The correlations used were obtained with the data from Carver [I]. The first group of rows contains R_{xx}, R_{xy}, I, and information about x_6. The second group

of rows contains S_{xx}, $S_{xx}R_{xx}^{-1}R_{xy}$, $S_{xx}R_{xx}^{-1}$ and information of x_i. The third group of rows contains the different sets of standard regression coefficients, the squares of the corresponding alienation coefficients, the diagonal terms of each corresponding inverse, the coefficients of alienation, the standard deviation of residuals, and the multiple correlation coefficients. The main results in the third group of rows are computed, as indicated in Chapter 13, by column-by-column multiplication of the results in the second rows.

18.9 Canonical correlation. Various methods discussed in earlier chapters are applied in this section to the computation of canonical correlation. The canonical correlation coefficient is the highest correlation coefficient that can be found in correlating the linear forms $u = xa$ and $v = yb$ by assigning suitable values to a and b. Here x is the matrix of the deviations of the observations of the first set of variables, y is the matrix of the deviations of the observations of the second set of variables, a is the (unknown) column matrix of the coefficients of x, b is the (unknown) column matrix of the coefficients of y, so that uv, $u'u$, and $v'v$ are scalar. It follows that

$$(1) \qquad \rho = \frac{v'u}{\sqrt{(u'u)(v'v)}} = \frac{b'y'xa}{\sqrt{(a'x'xa)(b'y'yb)}} = \frac{a'x'yb}{\sqrt{(a'x'xa)(b'y'yb)}}.$$

We simplify (1) by changing the scale so that the denominator is unity. In this case we have

$$(2) \qquad \rho = b'y'xa = a'x'yb$$

with the additional conditions

$$(3) \qquad a'x'xa = 1 \qquad \text{and} \qquad b'y'yb = 1.$$

It is desired to find the values of a and b that will minimize the values of ρ. We make use of Lagrange multipliers and get

$$(4) \qquad \rho = b'y'xa + \frac{c}{2}(1 - a'x'xa) + \frac{d}{2}(1 - b'y'yb).$$

Differentiation with respect to a [J] gives

$$(5) \qquad x'yb = cx'xa.$$

Premultiplication by a' gives

$$(6) \qquad a'x'yb = ca'x'xa = c$$

so that

$$c = \rho.$$

Similarly differentiation of (4) with respect to b gives

$$y'xa = dy'yb \tag{7}$$

and premultiplication by b' gives

$$d = \rho.$$

It is possible to use any of the deviates of section 18.6 in the analysis. If the small standard deviates are used, we have the conventional equations

$$\begin{aligned} \rho R_{xx}a &= R_{xy}b \\ R'_{xy}a &= \rho R_{yy}b, \end{aligned} \tag{8}$$

where the R's are matrices of correlation coefficients. These are the equations that have been used by previous writers [K].

There is no restriction in general in indicating that the R_{xx} matrix is the one that has the higher order if R_{yy} is of higher order than R_{xx}. In this case we get the simpler results if we eliminate the a, rather than the b, from (3). We obtain at once

$$(\rho^2 R_{yy} - R'_{xy}R_{xx}^{-1}R_{xy})b = 0$$

whence

$$(\rho^2 I - R_{yy}^{-1}R'_{xy}R_{xx}^{-1}R_{xy})b = 0. \tag{9}$$

We are confronted first with the problem of computing

$$L = R_{yy}^{-1}R'_{xy}R_{xx}^{-1}R_{xy}$$

and then solving the characteristic equation. It is conventional to compute $R_{yy}^{-1}R'_{xy}$ and $R_{xx}^{-1}R_{xy}$ and multiply the results. An easier method calls for the calculation of $R'_{xy}R_{xx}^{-1}R_{xy}$ and then the solution of the equation $R_{yy}(\ ) = R'_{xy}R_{xx}^{-1}R_{xy}$. The first of these is easily accomplished with the square root method, since it is necessary only to multiply $S_{xx}R_{xx}^{-1}R_{xy}$ by columns to get $R'_{xy}R_{xx}^{-1}R_{xy}$.

The method is illustrated in Table 18.9a, where the problem is based on correlations between two parts of the Thorndike Intelligence examination, an illustration previously used by Lorge [L]. The process leads to the calculation of the matrix L.

The next step proceeds with the calculation of the characteristic equation of L with the method of determinants by the method of section 15.3. The sums of the principal minors are indicated in the last column of the table, and the synthetic solution showing that $\rho^2 = 0.903$ is the highest value of ρ is shown at the bottom. The value of the

TABLE 18.9*a*

CORRELATIONS BETWEEN THE PARTS OF TWO FORMS OF THE THORNDIKE INTELLIGENCE EXAMINATION

Form	*A*			*B*			
Part	1	2	3	1	2	3	
R_{xx}	1.0000 * *	0.7830 1.0000 *	0.7852 0.8393 1.0000	0.8986 0.7961 0.7683	0.7841 0.8543 0.8226	0.8217 0.8254 0.8588	R_{xy}
S_{xx}	1.0000	0.7830 0.6220	0.7852 0.3609 0.5032	0.8986 0.1487 0.0180	0.7841 0.3864 0.1341	0.8217 0.2926 0.2146	$S_{xx}R_{xx}^{-1}R_{xy}$
R_{yy}	1.0000 * *	0.8235 1.0000 *	0.7912 0.8315 1.0000	0.8299 0.7645 0.7858	0.7645 0.7821 0.7861	0.7858 0.7861 0.8069	$R'_{xy}R_{xx}^{-1}R_{xy}$
S_{yy}	1.0000	0.8235 0.5673	0.7912 0.3172 0.5229	0.8299 0.1429 0.1604	0.7645 0.2689 0.1835	0.7858 0.2450 0.2055	
	1.0000			0.5210 0.0804	0.2581 0.2778	0.3002 0.2121	L 1.1918
		1.0000		0.3067	0.3509	0.3930	
			1.0000	(0.5210)	0.1240 0.1037	0.0864 0.1127	0.2714
				(0.2778)		0.0347	
						0.0096	0.0096
				0.3820 −0.0804 −0.3067	−0.2581 0.6252 −0.3509	−0.3002 −0.2121 0.5100	$\rho^2 I - L$
					0.2181 −0.2132	−0.1052 0.1027 −0.0001	
	1.1910 2.1989 5.3398	0.6907 2.1225 2.3108	0.5948 2.1097	1.1117 2.3001 5.9128	0.4823 2.2293 2.4316	1.0000 2.2806	
u	0.5155	0.2989	0.2574	0.4572	0.1983	0.4113	v

λ^3	λ^2	λ	λ^0	
1.0000	1.1918	0.2714	−0.0096	0.9030
	0.9030	−0.2608	0.0096	
1.0000	−0.2888	0.0106		

The solution of the resulting quadratic is:

$$\lambda = 0.2459, \quad \lambda = 0.0429 \quad \text{so that}$$
$$\rho = 0.9503, \quad \rho = 0.4958, \quad \rho = 0.2071.$$

matrix $\rho^2 I - L$ of (9) is then computed and the equations solved. It is shown that the b values are in the ratio

(10) $$1.1117 : 0.4823 : 1.0000.$$

A back solution obtains corresponding ratios for the a's, and, following Lorge, changes in scale are made to get proportional values of a and b in more suitable form.

An alternative derivation of the characteristic equation of L and the determination of its characteristic vector, following the method of section 15.5, is given in Table 18.9*b*, although some of the values are

TABLE 18.9*b*

LORGE PROBLEM—FRAME METHOD

0.5210	0.2581	0.3002	0.5210		0.8154*I*
0.0804	0.2778	0.2121	0.2778		
0.3067	0.3509	0.3930	0.3930		
			1.1918	1.1918	
−0.6708	0.0804	0.3067	−0.23666422		0.9030
0.2581	−0.9140	0.3509	−0.15873207		
0.3002	0.2121	−0.7988	−0.14743117		
−0.1125	−0.6215	−0.1412	−0.54282746	−0.27141373	
0.0347	0.0335	−0.0570	0.0096		1
0.0039	0.1127	−0.1037	0.0096		
−0.0287	−0.0864	0.1240	0.0096		
0.0099	0.0598	−0.0367	0.0288	0.0096	
0	0	0	0		
0	0	0	0		
0	0	0	0		
0	0	0	0	0	
0.2444	0.1661	0.2200	$\rho^2 = \lambda = 0.903$		
0.2370	0.1028	0.2132			
0.2424	0.1051	0.2181			
1.1109	0.4823	1.0000			
1.1116	0.4822	1.0000			
1.1114	0.4819	1.0000			

rounded off. This method results in substantially the same characteristic vector as that of (10).

18.10 The accumulation of errors. It is conventional to consider the sampling error in discussing a statistic. There are other errors that may accumulate and amount to more than the sampling error. In particular the measuring error, the recording error, and the computational error may amount to more than the sampling error for large samples. To be sure we may expect that these errors will be reduced by the principle of the combination of independent errors, but there is no guarantee that the measurement of errors is random. It may not be wise to write the value of a mean of 1000 heights as 67.32 ± 0.05 inches when the original measures were measured to the nearest inch. It could be, if the errors were all in the same direction, that the mean is in error by as much as 0.50 as a result of inadequate measurement or incomplete recording. In such a case it is not wise to emphasize the importance of the sampling error when there may be other errors, not even considered, up to ten times as large. Of course the measurement errors are usually random, but the possibility exists that they are not random. The investigator should study the nature of his measurements and use theory appropriate to his experimental knowledge rather than to assume randomness without discussion.

With modern computational machinery it is possible to avoid most grouping errors completely by recording and working with the precise measurements [M]. Usually the measurements can be made with such a degree of precision that the resulting grouping errors are trivial.

18.11 Conclusion. It is possible to expand the subject matter of this chapter into volumes, but it has seemed appropriate to indicate only a few representative applications.

REFERENCES

A. R. A. Fisher, *Statistical Methods for Research Workers*, Tenth Edition, Oliver and Boyd, London, 1946. See section 24.

B. P. R. Rider, *Statistical Methods*, John Wiley and Sons, New York, 1939. See section 43.

C. G. W. Snedecor, *Statistical Methods*, Fourth Edition, Iowa State College Press, 1946. See p. 254.

D. C. H. Goulden, *Methods of Statistical Analysis*, John Wiley and Sons, New York, 1939. See p. 131.

E. P. S. Dwyer, "A matrix presentation of least squares theory, etc.," *Annals of Mathematical Statistics*, **15** (1944), pp. 82–89.

F. See references in [F] of Chapter 6 to use of cracovians.

G. D. B. Duncan and J. F. Kenney, *On the Solution of Normal Equations and Related Topics*, Edwards Brothers, Ann Arbor, Mich., 1946.

H. P. S. Dwyer, "The square root method and its use in correlation and regression," *Journal of the American Statistical Association*, **40**, 502 (1945).

I. H. C. Carver, *Anthropometric Data*, Edwards Brothers, Ann Arbor, Mich., 1941.

J. P. S. Dwyer and M. S. Macphail, "Symbolic matrix derivatives," *Annals of Mathematical Statistics*, **19**, 517–534 (1948).

K. 1. H. Hotelling, "Relations between two sets of variates," *Biometrika*, **28**, 321–377 (1936).

2. F. V. Waugh, "Regression between sets of variates," *Econometrika*, **10** (1942), pp. 290–310.

L. I. Lorge, "The computation of the Hotelling canonical correlation," *Proceedings of the Educational Research Forum*, Aug. 26–31, 1940, International Business Machines Corporation, New York, pp. 68–74.

M. P. S. Dwyer, "Grouping methods," *Annals of Mathematical Statistics*, **13**, 138–155 (1942).

EXERCISES

No list of exercises is provided for this chapter since many problems are available in textbooks on statistics to those readers who have a special interest in this field.

CHAPTER 19

Application to Non-linear Problems Concluding Remarks

19.1 Applications to non-linear problems. Many of the techniques described in this treatise can be used in connection with non-linear problems: (*a*) by using the linear formulation as an approximation to the more general problem, (*b*) by making transformations that reduce the non-linear problem to linear form, and (*c*) by using interpolation formulas that are composed, essentially, of linear operations.

19.2 The linear formulation as an approximation. In conventional least squares problems the linear form is postulated as the skeleton on which the solution is based. As is commonly understood, this linear form is not necessarily the best postulated form, but it serves as a reasonable and practical mathematical model that can be used as a basis of reducing the data. Sometimes such linear approximation is made when a better mathematical model is unknown and sometimes it is made when a better model, though known, does not lend itself to suitable calculational techniques. As knowledge advances and as more suitable calculational techniques are made available, linear problems will probably not dominate the field as they do now. For some time at least, approximate answers will be obtained through the use of linear models.

19.3 The use of substitutions. There are many problems whose mathematical models call for exponential or power laws rather than linear form. Many of them can be reduced to linear form by transformations. For example, a problem involving the fitting of a parabola $Y = b_0 + b_1X + b_2X^2$ can be reduced to linear form by substituting X_2 for X^2 so that the mathematical model is

$$Y = b_0 + b_1X_1 + b_2X_2. \tag{1}$$

Such reduction is proper since the general theory of least squares does not demand that X_1 and X_2 be independent. They are, in fact, usually correlated. Of course the significance of b_2 can be tested, and, if b_2 is shown not to differ significantly from zero, then the linear model is probably not too unsatisfactory.

TABLE 19.3*a*

PARABOLIC REGRESSION—METHOD OF DETERMINANTS

i	$X_1 = X$	$X_2 = X^2$	Y	
1	X_{11}	X_{21}	Y_1	
2	X_{12}	X_{22}	Y_2	
3	X_{13}	X_{23}	Y_3	
. . .	. . . . .	. . . . .	. . . . .	
N	X_{1N}	X_{2N}	Y_N	
N	$\Sigma X_1 = \Sigma X$	$\Sigma X_2 = \Sigma X^2$	ΣY	
	$\Sigma X_1^2 = \Sigma X^2$	$\Sigma X_1 X_2 = \Sigma X^3$	$\Sigma X_1 Y = \Sigma XY$	
		$\Sigma X_2^2 = \Sigma X^4$	$\Sigma X_2 Y = \Sigma X^2 Y$	
			ΣY^2	
	L_{11}	L_{12}	L_{1y}	
		L_{22}	L_{2y}	
			L_{yy}	
		$d_{22 \cdot 1}$	$d_{2y \cdot 1}$	
			$d_{yy \cdot 1}$	
$b_{0y \cdot 1}$	b_{1y}		ΣE_1^2	DF
$b_{0y \cdot 12}$	$b_{1y \cdot 2}$	$b_{2y \cdot 1}$	ΣE_2^2	DF
			Diff.	DF

An illustration of the application of the method of determinants to a problem of this sort used by Rider [A] is given in Table 19.3*b*. The problem is fitting a parabola through the points (0,1), (1,3), (3,2), (6,5), and (8,4). These coordinates are exact numbers and are not subject to error.

The transformation is fairly trivial for the parabola, but many more complex expressions can be reduced to linear form by suitable transformations. The reader may be interested in referring to an article by Huntington [B], where several least squares solutions are presented.

TABLE 19.3b

PARABOLIC REGRESSION—METHOD OF DETERMINANTS
ILLUSTRATION

i	X	X^2	Y	
1	0	0	1	
2	1	1	3	
3	3	9	2	
4	6	36	5	
5	8	64	4	
5	18	110	15	
	110	756	71	
		5474	457	
			55	
	226	1800	85	
		15270	635	
			50	
		42204	−1898	
			815	
1.6460	0.3761		3.606	1
1.3458	0.7345	−0.0450	3.229	2
			0.377	1

19.4 The use of interpolation. A third use of linear methods in non-linear problems utilizes linear interpolation, which is based on the fact that curves may be approximated by a series of straight lines. Without going into a detailed discussion of an extensive subject, we may point out interpolation techniques that make possible a good calculational design when a desk machine is used.

A common formula for direct linear interpolation when the values of $f(a)$ and $f(a + 1)$ are tabulated is

$$f(x + \theta) = f(a) + \theta[f(a + 1) - f(a)]. \tag{1}$$

This formula can be put in the form

$$f(x + \theta) = \theta f(a + 1) + (1 - \theta)f(a) \tag{2}$$

which is an operational unit and can be performed directly from the successive entries of the table.

Conventionally the inverse of (1) is used in the form

$$\theta = \frac{f(a+\theta) - f(a)}{f(a+1) - f(a)} \tag{3}$$

and the result is added to a to obtain the value of x. If we perform the addition first, we get

$$a + \theta = a + \frac{f(a+\theta) - f(a)}{f(a+1) - f(a)}$$

so that

$$x = \frac{af(a+1) - (a+1)f(a) + f(x)}{f(a+1) - f(a)}. \tag{4}$$

This too is an operational unit of type U_9.

A computing form is sometimes useful, such as

x	$f(x)$
$a+1$	$f(a+1)$
$x = ?$	$f(x)$
a	$f(a)$
	$f(a+1) - f(a)$

The available values of x and $f(x)$ are tabulated and the difference $f(a+1) - f(a)$ is listed. The formula (4) is then translated to a simple ritual.

It is possible to apply this general scheme to a slightly more general problem. Suppose that the values of x, for which the values of $f(x)$ are available, are a and b. Then the basic formula is

$$\frac{f(x) - f(a)}{f(b) - f(a)} = \frac{x - a}{b - a}. \tag{5}$$

We solve for $f(x)$ and reduce to the form

$$f(x) = \frac{bf(a) - af(b) + x[f(b) - f(a)]}{b - a}. \tag{6}$$

We solve for x and get

$$x = \frac{af(b) - bf(a) + (b-a)f(x)}{f(b) - f(a)}. \tag{7}$$

TABLE 19.3b

PARABOLIC REGRESSION—METHOD OF DETERMINANTS
ILLUSTRATION

i	X	X^2	Y	
1	0	0	1	
2	1	1	3	
3	3	9	2	
4	6	36	5	
5	8	64	4	
5	18	110	15	
	110	756	71	
		5474	457	
			55	
	226	1800	85	
		15270	635	
			50	
		42204	−1898	
			815	
1.6460	0.3761		3.606	1
1.3458	0.7345	−0.0450	3.229	2
			0.377	1

19.4 The use of interpolation. A third use of linear methods in non-linear problems utilizes linear interpolation, which is based on the fact that curves may be approximated by a series of straight lines. Without going into a detailed discussion of an extensive subject, we may point out interpolation techniques that make possible a good calculational design when a desk machine is used.

A common formula for direct linear interpolation when the values of $f(a)$ and $f(a + 1)$ are tabulated is

$$f(x + \theta) = f(a) + \theta[f(a + 1) - f(a)]. \tag{1}$$

This formula can be put in the form

$$f(x + \theta) = \theta f(a + 1) + (1 - \theta)f(a) \tag{2}$$

which is an operational unit and can be performed directly from the successive entries of the table.

Conventionally the inverse of (1) is used in the form

$$\theta = \frac{f(a+\theta) - f(a)}{f(a+1) - f(a)} \tag{3}$$

and the result is added to a to obtain the value of x. If we perform the addition first, we get

$$a + \theta = a + \frac{f(a+\theta) - f(a)}{f(a+1) - f(a)}$$

so that

$$x = \frac{af(a+1) - (a+1)f(a) + f(x)}{f(a+1) - f(a)}. \tag{4}$$

This too is an operational unit of type U_9.

A computing form is sometimes useful, such as

x	$f(x)$
$a+1$	$f(a+1)$
$x = ?$	$f(x)$
a	$f(a)$
	$f(a+1) - f(a)$

The available values of x and $f(x)$ are tabulated and the difference $f(a+1) - f(a)$ is listed. The formula (4) is then translated to a simple ritual.

It is possible to apply this general scheme to a slightly more general problem. Suppose that the values of x, for which the values of $f(x)$ are available, are a and b. Then the basic formula is

$$\frac{f(x) - f(a)}{f(b) - f(a)} = \frac{x - a}{b - a}. \tag{5}$$

We solve for $f(x)$ and reduce to the form

$$f(x) = \frac{bf(a) - af(b) + x[f(b) - f(a)]}{b - a}. \tag{6}$$

We solve for x and get

$$x = \frac{af(b) - bf(a) + (b-a)f(x)}{f(b) - f(a)}. \tag{7}$$

The computing form

x	$f(x)$
b	$f(b)$
x	$f(x)$
a	$f(a)$
$b - a$	$f(b) - f(a)$

is adaptable to the use of either (6) or (7).

Now (4) is a special case of (7), with $b = a + 1$, and (2) is a special case of (6), which may be written in the form

$$f(x) = \frac{(x - a)f(b) + (b - x)f(a)}{b - a}. \tag{8}$$

Thus (8) is a direct generalization of (2).

A more general formula, however, can be derived that has the direct and inverse interpolation formulas as special cases and permits a direct change from one function to another. Suppose the one function is $f(x)$ and the second function is $g(x)$ and that they are tabulated for the values of a and b. The basic interpolation identity is

$$\frac{f(x) - f(a)}{f(b) - f(a)} = \frac{g(x) - g(a)}{g(b) - g(a)}.$$

We solve this for $g(x)$ to get

$$g(x) = \frac{g(a)f(b) - g(b)f(a) + [g(b) - g(a)]f(x)}{f(b) - f(a)}. \tag{9}$$

Now (9) is the equivalent of (6) when $f(x) = x$, and it becomes (7) when $g(x) = x$. The computing form

$f(x)$	$g(x)$
$f(b)$	$g(b)$
$f(x)$	$g(x)$
$f(a)$	$g(a)$
$f(b) - f(a)$	$b - a$

is good for both cases.

Also linear methods can be used in connection with interpolation higher order. For example, the parabola

$$y = b_0 + b_1x + b_2x^2$$

may be used as the interpolating curve where the corresponding values of x and y are known for three points. In this case we have

$$\begin{aligned} b_0 + b_1x_1 + b_2x_1^2 &= y_1 \\ b_0 + b_1x_2 + b_2x_2^2 &= y_2 \\ b_0 + b_1x_3 + b_2x_3^2 &= y_3 \\ b_0 + b_1x + b_2x^2 &= y \end{aligned} \tag{10}$$

and the condition is that

$$\begin{vmatrix} 1 & x_1 & x_1^2 & y_1 \\ 1 & x_2 & x_2^2 & y_2 \\ 1 & x_3 & x_3^2 & y_3 \\ 1 & x & x^2 & y \end{vmatrix} = 0. \tag{11}$$

To find the value of y on the interpolating curve for any specified value of x, insert the values of x_i and y_i and evaluate Δ by some method such as that of Table 11.3*a*.

It is possible to obtain a number of the conventional interpolation formulas for (11). We get Newton's formula by placing $x_1 = 0$, $x_2 = 1$, $x_3 = 2$ and reduce to obtain the formula [C]

$$y = y_0 + x\,\Delta y_0 + \frac{x(x-1)}{2}\Delta^2 y_0$$

where $\Delta y_0 = y_1 - y_0$, and $\Delta^2 y_0 = y_2 - 2y_1 + y_0$. If we put $x_1 = -1$, $x_2 = 0$, and $x_3 = 1$ in (11), we get

$$y = y_0 + x\,\frac{\Delta y_0 + \Delta y_1}{2} + \frac{x^2}{2}\Delta^2 y_0$$

which is the type commonly known as Stirling's interpolation formula.

Now (11) may be expanded to get a variety of interpolation formulas. Among the most useful of these, for those who have access to machines and appropriate tables, are the Lagrange formulas.

To develop (11) and exhibit y as a linear function of y_1, y_2, and y_3, we simplify it to get

$$y = \frac{(x-x_2)(x-x_3)}{(x_1-x_2)(x_1-x_3)}y_1 + \frac{(x-x_1)(x-x_3)}{(x_2-x_1)(x_2-x_3)}y_2 + \frac{(x-x_1)(x-x_2)}{(x_3-x_1)(x_3-x_2)}y_3 \tag{12}$$

We see that the values of y for any specific x can be computed with the use of

$$y = L_1y_1 + L_2y_2 + L_3y_3, \tag{13}$$

where the L_i are the Lagrangian interpolation coefficients.

The equation (12) does not need any of the earlier justification since it obviously contains the points (x_1,y_1), (x_2,y_2) and (x_3,y_3) and is of the second order in x. In a similar way it can be seen by inspection that

$$\begin{aligned} y = {} & \frac{(x - x_2)(x - x_3)(x - x_4)}{(x_1 - x_2)(x_1 - x_3)(x_1 - x_4)} y_1 \\ & + \frac{(x - x_1)(x - x_3)(x - x_4)}{(x_2 - x_1)(x_2 - x_3)(x_2 - x_4)} y_2 \\ & + \frac{(x - x_1)(x - x_2)(x - x_4)}{(x_3 - x_1)(x_3 - x_2)(x_3 - x_4)} y_3 \\ & + \frac{(x - x_1)(x - x_2)(x - x_3)}{(x_4 - x_1)(x_4 - x_2)(x_4 - x_3)} y_4 \\ = {} & L_1y_1 + L_2y_2 + L_3y_3 + L_4y_4 \end{aligned} \tag{14}$$

is the desired Lagrangian formula of the third order in x, since it passes through the points (x_1,y_1), (x_2y_2), (x_3y_3), and (x_4,y_4). The Lagrangian interpolating polynomials of higher order can be written similarly by inspection.

Effective use of Lagrangian polynomials calls for the previous calculation of the L_i. In case the ordinates are equally spaced the denominators of the L_i are constant and independent of x. Thus the values of L_i in (14) become, if the ordinates are spaced at unit distances,

$$\begin{aligned} L_1 &= \frac{(x - x_2)(x - x_3)(x - x_4)}{-6} \\ L_2 &= \frac{(x - x_1)(x - x_3)(x - x_4)}{2} \\ L_3 &= \frac{(x - x_1)(x - x_2)(x - x_4)}{-2} \\ L_4 &= \frac{(x - x_1)(x - x_2)(x - x_3)}{6}. \end{aligned} \tag{15}$$

It is frequently feasible to select the ordinates so as to center the value of x. Thus we can place $x = -3/2$, $x = -1/2$, $x = 1/2$, $x = 3/2$. If we wish the value when $x = 0$, we have

$$L_1 = -\tfrac{1}{16}, \qquad L_2 = \tfrac{9}{16}, \qquad L_3 = \tfrac{9}{16}, \qquad L_4 = -\tfrac{1}{16}.$$

It follows that (14) becomes, under these considerations,

$$y = \frac{-y_1 + 9y_2 + 9y_3 - y_4}{16}$$

$$= -0.0625y_1 + 0.5625y_2 + 0.5625y_3 - 0.0625y_4.$$

The computational technique calls for the cumulative multiplication of the successive ordinates by the values -1, 9, 9, -1 and the division by 16 or by multiplication by the values -0.0625, 0.5625, 0.5625, -0.0625.

This technique could be applied in using third-order interpolation to obtain the values of a function in one-half units that are tabulated in units. Suppose, for example, that it is desired to tabulate $\sqrt{N}$ for half intervals when the approximate values of $\sqrt{N}$ are known for intervals. It is possible to write the known values in alternating rows and fill in the others as computed. The process is illustrated in Table 19.4*a*. The

TABLE 19.4*a*

ILLUSTRATION OF LAGRANGIAN INTERPOLATION

N	$\sqrt{N}$	
...	....	
4970	70.49827	
4975		
4980	70.5691	
4985	70.605	-1
4990	70.6399	0
4995	70.675	9
5000	70.7107	←
5005	70.746	9
5010	70.7814	0
5015	70.817	-1
5020	70.8520	
5025	70.887	
5030	70.9225	
5035		
5040	70.9930	
...	....	

calculation may be facilitated with an auxiliary strip of paper marked $-1, 0, 9, \leftarrow, 9, 0, -1$ in successive rows. This could be placed adjacent to the values of y so as to bring together the values to be multiplied. The arrow head then indicates the position for the result.

Tables of Lagrangian coefficients are available [D] that present extensive 3-point and 4-point formulas, and less extensive 5, 6, 7, 8, 9, 10, 11 point formulas. Before using these tables, we should spend some time with the introductory material so that we shall be able to translate our problem into the language of the tables. The coefficients for the illustration above are found on page 203 and again on page 371.

19.5 Concluding remarks. This is not all there is to say about linear computations. The term linear problems, interpreted broadly, includes much more than the problems discussed in this book. An attempt has been made to present a fairly extensive and detailed discussion of such basic linear problems as the solution of simultaneous linear equations, the evaluation of determinants, the calculation of the adjoint and the inverse of a matrix, and the solution of problems involving the characteristic equation by direct methods and by a desk calculating machine. Some mention has been made of iterative methods and enlargement methods but, in the main, they have not received as much attention as direct elimination or condensation methods.

This book is designed primarily for the individual worker and hence emphasizes methods and techniques that are especially applicable to the modern desk calculator, although many of the methods and suggestions are applicable to other computational devices. It is the thesis of the author, however, that a greatly increasing amount of calculational work will be carried on by operation with a desk calculator as more and more research workers in more and more fields become aware of the necessity, feasibility, and practicability of the reduction of data through mathematical and statistical devices. This does not say that there will not be enormous expansion in mass calculations with much more complex machines, such as punched card calculators or electronic digital computers, which are capable of utilizing more complex operational units than are possible with a desk computer. These machines, because they are designed for complex operations, cannot be used most efficiently in performing simple operations. This task has to be carried on by the simpler machines, which include the desk calculator, as well as the punched card sorter and tabulator.

The methods have now been perfected so that it is feasible to solve problems involving up to fifteen or twenty variables with the desk computer. Though we cannot draw hard and fast rules that are applicable to every situation, it appears that a research worker can handle

problems of this size quite efficiently in his own office without resorting to machines of huge capacity in some distant computing center. There is every reason to think that a major portion of the studies to be made in the future will, as they have in the past, contain less than twenty variables. In most cases in which very many variables have been used, it has been demonstrated that a few of them give essentially the basic information of all of them. Preliminary studies should reveal which are the essential variables and then the grand experiment should be designed on the basis of those variables that give some real promise and not on the "drag net policy" of throwing in every variable we can think of and seeing if it makes a contribution. We may argue that a drag net policy is a good policy since only in this way can we make sure that we have everything possible. This argument seems fair enough, but if, after preliminary reduction of data, inspection shows that many of these variables are making no contribution, they can immediately be discarded and should not enter the linear statement of the problem at all. Very rarely are the number of variables that might make a contribution greater than twenty. Even if the number were greater than twenty, the mere effort of collecting all the data is so enormous that it would take a whole research organization, and not just a research worker or two, to collect them and process them. Time spent in collecting and reducing the data that serve as the basis of a linear problem involving a hundred variables, for example, is enormous when compared with the actual time necessary to solve the equations. It begins to be apparent that when the electronic digital computers are perfected so that they can solve linear problems involving many variables easily, as they give every promise of doing, they will not be solving the majority of problems of this type, at least not those based on experimental evidence.

There is some evidence, on the basis of the use of complex computing machines now in operation, that the machines will be used more and more in connection with the complex operations of evaluating functions or computing tables and less and less in connection with the reduction of experimental data and the solution of linear problems resulting therefrom.

It appears then that the desk machine (and it must be remembered that it will probably be perfected too, perhaps first with the addition of electronic storage counters) is capable of handling the great majority of the primary linear problems after the preliminary reduction of the data is made with it, or with more primitive machines. The occasional problem involving very many variables will probably be handled by the big electronic digital computer. It now appears that more likely rivals

for the desk computers in solving linear problems, since they are, under certain conditions, efficient in handling linear problems when there are six to twenty variables, are

(*a*) Punched card computing machines.
(*b*) Analogue equation solvers.
(*c*) Small-scale electronic digital computers.

Some of the special conditions under which one or more of these rival computers is especially appropriate are:

(*a*) When many sets of simultaneous equations are to be solved. For instance, we might wish to solve 250 sets of equations, each consisting of five equations in five unknowns.
(*b*) When the sets of equations take on a form that is especially adapted to the particular machine. For instance, if the diagonal terms are large with respect to the non-diagonal terms, it is possible to combine the use of analogue computers and the iterative process and to obtain approximate answers in a very short time.

The methods of this book have been worked out particularly for desk machines and, as such, feature elimination methods. Some of these methods can be adapted to punched card machines, but usually with some loss in flexibility. For example, a detailed and not a compact form of the method of single division should be used. Thus the calculation would lead, from the values of a_{ij} to the values of $g_{ij \cdot 1}$, etc.

This book does not feature iterative methods because the direct methods, either exact or approximate, seem satisfactory. Iterative methods may be more useful in connection with punched card machines or in connection with large-scale electronic computers.

It is indicated that some large-scale digital computers will be capable of matrix multiplication and perhaps matrix inversion. In such a case the matrix statement of the problems appearing in Chapters 13 to 17 can be directly transferred to a calculational technique.

The research worker should re-examine all his calculational techniques when an important addition has been made to his mechanical equipment to see if a new design may not enable him to use the full capacities of the machine to better advantage.

REFERENCES

A. P. R. Rider, *Statistical Methods*, John Wiley and Sons, New York, 1939, p. 42.

B. E. V. Huntington, "Curve fitting by the method of least squares and the method of moments," *Handbook of Mathematical Statistics*, Houghton Mifflin Co., Boston, 1924, pp. 62–70.

C. J. W. Glover, "Interpolation, summation, and graduation," *Handbook of Mathematical Statistics*, Houghton Mifflin Co., Boston, 1924, pp. 34–61. See pp. 36–38.

D. Mathematical Tables Project, *Tables of Lagrangian Interpolation Coefficients*, Columbia University Press, New York, 1944.

EXERCISES

No list of exercises is presented for this chapter since many illustrations are presumably well known to any reader who may be interested in the special techniques covered.

Author References

Index