Advanced Calculus with Applications in Statistics

Second Edition
Revised and Expanded

André I. Khuri

University of Florida
Gainesville, Florida

A JOHN WILEY & SONS, INC., PUBLICATION

Advanced Calculus with Applications in Statistics

Second Edition

Advanced Calculus with Applications in Statistics

Second Edition
Revised and Expanded

André I. Khuri

University of Florida
Gainesville, Florida

A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data

Khuri, André I., 1940-
Advanced calculus with applications in statistics / André I. Khuri. -- 2nd ed. rev. and expended.
p. cm. -- (Wiley series in probability and statistics)
Includes bibliographical references and index.
ISBN 0-471-39104-2 (cloth : alk. paper)
1. Calculus. 2. Mathematical statistics. I. Title. II. Series.

QA303.2.K48 2003
515--dc21 2002068986

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

To Ronnie, Marcus, and Roxanne
and
In memory of my sister Ninette

Contents

Preface

This edition provides a rather substantial addition to the material covered in the first edition. The principal difference is the inclusion of three new chapters, Chapters 10, 11, and 12, in addition to an appendix of solutions to exercises.

Chapter 10 covers orthogonal polynomials, such as Legendre, Chebyshev, Jacobi, Laguerre, and Hermite polynomials, and discusses their applications in statistics. Chapter 11 provides a thorough coverage of Fourier series. The presentation is done in such a way that a reader with no prior knowledge of Fourier series can have a clear understanding of the theory underlying the subject. Several applications of Fouries series in statistics are presented. Chapter 12 deals with approximation of Riemann integrals. It gives an exposition of methods for approximating integrals, including those that are multidimensional. Applications of some of these methods in statistics are discussed. This subject area has recently gained prominence in several fields of science and engineering, and, in particular, Bayesian statistics. The material should be helpful to readers who may be interested in pursuing further studies in this area.

A significant addition is the inclusion of a major appendix that gives detailed solutions to the vast majority of the exercises in Chapters 1–12. This supplement was prepared in response to numerous suggestions by users of the first edition. The solutions should also be helpful in getting a better understanding of the various topics covered in the book.

In addition to the aforementioned material, several new exercises were added to some of the chapters in the first edition. Chapter 1 was expanded by the inclusion of some basic topological concepts. Chapter 9 was modified to accommodate Chapter 10. The changes in the remaining chapters, 2 through 8, are very minor. The general bibliography was updated.

The choice of the new chapters was motivated by the evolution of the field of statistics and the growing needs of statisticians for mathematical tools beyond the realm of advanced calculus. This is certainly true in topics concerning approximation of integrals and distribution functions, stochastic

processes, time series analysis, and the modeling of periodic response functions, to name just a few.

The book is self-contained. It can be used as a text for a two-semester course in advanced calculus and introductory mathematical analysis. Chapters 1–7 may be covered in one semester, and Chapters 8–12 in the other semester. With its coverage of a wide variety of topics, the book can also serve as a reference for statisticians, and others, who need an adequate knowledge of mathematics, but do not have the time to wade through the myriad mathematics books. It is hoped that the inclusion of a separate section on applications in statistics in every chapter will provide a good motivation for learning the material in the book. This represents a continuation of the practice followed in the first edition.

As with the first edition, the book is intended as much for mathematicians as for statisticians. It can easily be turned into a pure mathematics book by simply omitting the section on applications in statistics in a given chapter. Mathematicians, however, may find the sections on applications in statistics to be quite useful, particularly to mathematics students seeking an interdisciplinary major. Such a major is becoming increasingly popular in many circles. In addition, several topics are included here that are not usually found in a typical advanced calculus book, such as approximation of functions and integrals, Fourier series, and orthogonal polynomials. The fields of mathematics and statistics are becoming increasingly intertwined, making any separation of the two unpropitious. The book represents a manifestation of the interdependence of the two fields.

The mathematics background needed for this edition is the same as for the first edition. For readers interested in statistical applications, a background in introductory mathematical statistics will be helpful, but not absolutely essential. The annotated bibliography in each chapter can be consulted for additional readings.

I am grateful to all those who provided comments and helpful suggestions concerning the first edition, and to my wife Ronnie for her help and support.

ANDRÉ I. KHURI

Gainesville, Florida

Preface to the First Edition

The most remarkable mathematical achievement of the seventeenth century was the invention of calculus by Isaac Newton (1642–1727) and Gottfried Wilhelm Leibniz (1646–1716). It has since played a significant role in all fields of science, serving as its principal quantitative language. There is hardly any scientific discipline that does not require a good knowledge of calculus. The field of statistics is no exception.

Advanced calculus has had a fundamental and seminal role in the development of the basic theory underlying statistical methodology. With the rapid growth of statistics as a discipline, particularly in the last three decades, knowledge of advanced calculus has become imperative for understanding the recent advances in this field. Students as well as research workers in statistics are expected to have a certain level of mathematical sophistication in order to cope with the intricacies necessitated by the emerging of new statistical methodologies.

This book has two purposes. The first is to provide beginning graduate students in statistics with the basic concepts of advanced calculus. A high percentage of these students have undergraduate training in disciplines other than mathematics with only two or three introductory calculus courses. They are, in general, not adequately prepared to pursue an advanced graduate degree in statistics. This book is designed to fill the gaps in their mathematical training and equip them with the advanced calculus tools needed in their graduate work. It can also provide the basic prerequisites for more advanced courses in mathematics.

One salient feature of this book is the inclusion of a complete section in each chapter describing applications in statistics of the material given in the chapter. Furthermore, a large segment of Chapter 8 is devoted to the important problem of optimization in statistics. The purpose of these applications is to help motivate the learning of advanced calculus by showing its relevance in the field of statistics. There are many advanced calculus books designed for engineers or business majors, but there are none for statistics

majors. This is the first advanced calculus book to emphasize applications in statistics.

The scope of this book is not limited to serving the needs of statistics graduate students. Practicing statisticians can use it to sharpen their mathematical skills, or they may want to keep it as a handy reference for their research work. These individuals may be interested in the last three chapters, particularly Chapters 8 and 9, which include a large number of citations of statistical papers.

The second purpose of the book concerns mathematics majors. The book's thorough and rigorous coverage of advanced calculus makes it quite suitable as a text for juniors or seniors. Chapters 1 through 7 can be used for this purpose. The instructor may choose to omit the last section in each chapter, which pertains to statistical applications. Students may benefit, however, from the exposure to these additional applications. This is particularly true given that the trend today is to allow the undergraduate student to have a major in mathematics with a minor in some other discipline. In this respect, the book can be particularly useful to those mathematics students who may be interested in a minor in statistics.

Other features of this book include a detailed coverage of optimization techniques and their applications in statistics (Chapter 8), and an introduction to approximation theory (Chapter 9). In addition, an annotated bibliography is given at the end of each chapter. This bibliography can help direct the interested reader to other sources in mathematics and statistics that are relevant to the material in a given chapter. A general bibliography is provided at the end of the book. There are also many examples and exercises in mathematics and statistics in every chapter. The exercises are classified by discipline (mathematics and statistics) for the benefit of the student and the instructor.

The reader is assumed to have a mathematical background that is usually obtained in the freshman–sophomore calculus sequence. A prerequisite for understanding the statistical applications in the book is an introductory statistics course. Obviously, those not interested in such applications need not worry about this prerequisite. Readers who do not have any background in statistics, but are nevertheless interested in the application sections, can make use of the annotated bibliography in each chapter for additional reading.

The book contains nine chapters. Chapters 1–7 cover the main topics in advanced calculus, while chapters 8 and 9 include more specialized subject areas. More specifically, Chapter 1 introduces the basic elements of set theory. Chapter 2 presents some fundamental concepts concerning vector spaces and matrix algebra. The purpose of this chapter is to facilitate the understanding of the material in the remaining chapters, particularly, in Chapters 7 and 8. Chapter 3 discusses the concepts of limits and continuity of functions. The notion of differentiation is studied in Chapter 4. Chapter 5 covers the theory of infinite sequences and series. Integration of functions is

the theme of Chapter 6. Multidimensional calculus is introduced in Chapter 7. This chapter provides an extension of the concepts of limits, continuity, differentiation, and integration to functions of several variables (multivariable functions). Chapter 8 consists of two parts. The first part presents an overview of the various methods of optimization of multivariable functions whose optima cannot be obtained explicitly by standard advanced calculus techniques. The second part discusses a variety of topics of interest to statisticians. The common theme among these topics is optimization. Finally, Chapter 9 deals with the problem of approximation of continuous functions with polynomial and spline functions. This chapter is of interest to both mathematicians and statisticians and contains a wide variety of applications in statistics.

I am grateful to the University of Florida for granting me a sabbatical leave that made it possible for me to embark on the project of writing this book. I would also like to thank Professor Rocco Ballerini at the University of Florida for providing me with some of the exercises used in Chapters, 3, 4, 5, and 6.

ANDRÉ I. KHURI

Gainesville, Florida

CHAPTER 1

An Introduction to Set Theory

The origin of the modern theory of sets can be traced back to the Russian-born German mathematician Georg Cantor (1845–1918). This chapter introduces the basic elements of this theory.

1.1. THE CONCEPT OF A SET

A set is any collection of well-defined and distinguishable objects. These objects are called the elements, or members, of the set and are denoted by lowercase letters. Thus a set can be perceived as a collection of elements united into a single entity. Georg Cantor stressed this in the following words: "A set is a multitude conceived of by us as a one."

If x is an element of a set A, then this fact is denoted by writing $x \in A$. If, however, x is not an element of A, then we write $x \notin A$. Curly brackets are usually used to describe the contents of a set. For example, if a set A consists of the elements $x_1, x_2, \ldots, x_n$, then it can be represented as $A = \{x_1, x_2, \ldots, x_n\}$. In the event membership in a set is determined by the satisfaction of a certain property or a relationship, then the description of the same can be given within the curly brackets. For example, if A consists of all real numbers x such that $x^2 > 1$, then it can be expressed as $A = \{x | x^2 > 1\}$, where the bar | is used simply to mean "such that." The definition of sets in this manner is based on the axiom of abstraction, which states that given any property, there exists a set whose elements are just those entities having that property.

Definition 1.1.1. The set that contains no elements is called the empty set and is denoted by $\varnothing$. □

Definition 1.1.2. A set A is a subset of another set B, written symbolically as $A \subset B$, if every element of A is an element of B. If B contains at least one element that is not in A, then A is said to be a proper subset of B. □

Definition 1.1.3. A set A and a set B are equal if $A \subset B$ and $B \subset A$. Thus, every element of A is an element of B and vice versa. □

Definition 1.1.4. The set that contains all sets under consideration in a certain study is called the universal set and is denoted by Ω. □

1.2. SET OPERATIONS

There are two basic operations for sets that produce new sets from existing ones. They are the operations of union and intersection.

Definition 1.2.1. The union of two sets A and B, denoted by $A \cup B$, is the set of elements that belong to either A or B, that is,

$$A \cup B = \{x | x \in A \text{ or } x \in B\}. \qquad \square$$

This definition can be extended to more than two sets. For example, if $A_1, A_2, \ldots, A_n$ are n given sets, then their union, denoted by $\bigcup_{i=1}^{n} A_i$, is a set such that x is an element of it if and only if x belongs to at least one of the A_i $(i = 1, 2, \ldots, n)$.

Definition 1.2.2. The intersection of two sets A and B, denoted by $A \cap B$, is the set of elements that belong to both A and B. Thus

$$A \cap B = \{x | x \in A \text{ and } x \in B\}. \qquad \square$$

This definition can also be extended to more than two sets. As before, if $A_1, A_2, \ldots, A_n$ are n given sets, then their intersection, denoted by $\bigcap_{i=1}^{n} A_i$, is the set consisting of all elements that belong to all the A_i $(i = 1, 2, \ldots, n)$.

Definition 1.2.3. Two sets A and B are disjoint if their intersection is the empty set, that is, $A \cap B = \varnothing$. □

Definition 1.2.4. The complement of a set A, denoted by $\bar{A}$, is the set consisting of all elements in the universal set that do not belong to A. In other words, $x \in \bar{A}$ if and only if $x \notin A$.

The complement of A with respect to a set B is the set $B - A$ which consists of the elements of B that do not belong to A. This complement is called the relative complement of A with respect to B. □

From Definitions 1.1.1–1.1.4 and 1.2.1–1.2.4, the following results can be concluded:

RESULT 1.2.1. The empty set $\varnothing$ is a subset of every set. To show this, suppose that A is any set. If it is false that $\varnothing \subset A$, then there must be an

element in $\varnothing$ which is not in A. But this is not possible, since $\varnothing$ is empty. It is therefore true that $\varnothing \subset A$.

RESULT 1.2.2. The empty set $\varnothing$ is unique. To prove this, suppose that $\varnothing_1$ and $\varnothing_2$ are two empty sets. Then, by the previous result, $\varnothing_1 \subset \varnothing_2$ and $\varnothing_2 \geq \varnothing_1$. Hence, $\varnothing_1 = \varnothing_2$.

RESULT 1.2.3. The complement of $\varnothing$ is Ω. Vice versa, the complement of Ω is $\varnothing$.

RESULT 1.2.4. The complement of $\overline{A}$ is A.

RESULT 1.2.5. For any set A, $A \cup \overline{A} = \Omega$ and $A \cap \overline{A} = \varnothing$.

RESULT 1.2.6. $A - B = A - A \cap B$.

RESULT 1.2.7. $A \cup (B \cup C) = (A \cup B) \cup C$.

RESULT 1.2.8. $A \cap (B \cap C) = (A \cap B) \cap C$.

RESULT 1.2.9. $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.

RESULT 1.2.10. $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.

RESULT 1.2.11. $\overline{(A \cup B)} = \overline{A} \cap \overline{B}$. More generally, $\overline{\bigcup_{i=1}^{n} A_i} = \bigcap_{i=1}^{n} \overline{A}_i$.

RESULT 1.2.12. $\overline{(A \cap B)} = \overline{A} \cup \overline{B}$. More generally, $\overline{\bigcap_{i=1}^{n} A_i} = \bigcup_{i=1}^{n} \overline{A}_i$.

Definition 1.2.5. Let A and B be two sets. Their Cartesian product, denoted by $A \times B$, is the set of all ordered pairs (a, b) such that $a \in A$ and $b \in B$, that is,

$$A \times B = \{(a, b) \mid a \in A \text{ and } b \in B\}.$$

The word "ordered" means that if a and c are elements in A and b and d are elements in B, then $(a, b) = (c, d)$ if and only if $a = c$ and $b = d$. □

The preceding definition can be extended to more than two sets. For example, if $A_1, A_2, \ldots, A_n$ are n given sets, then their Cartesian product is denoted by $\times_{i=1}^{n} A_i$ and defined by

$$\bigtimes_{i=1}^{n} A_i = \{(a_1, a_2, \ldots, a_n) \mid a_i \in A_i,\ i = 1, 2, \ldots, n\}.$$

Here, $(a_1, a_2, \ldots, a_n)$, called an ordered n-tuple, represents a generalization of the ordered pair. In particular, if the A_i are equal to A for $i = 1, 2, \ldots, n$, then one writes A^n for $\times_{i=1}^{n} A$.

The following results can be easily verified:

RESULT 1.2.13. $A \times B = \varnothing$ if and only if $A = \varnothing$ or $B = \varnothing$.

RESULT 1.2.14. $(A \cup B) \times C = (A \times C) \cup (B \times C)$.

RESULT 1.2.15. $(A \cap B) \times C = (A \times C) \cap (B \times C)$.

RESULT 1.2.16. $(A \times B) \cap (C \times D) = (A \cap C) \times (B \cap D)$.

1.3. RELATIONS AND FUNCTIONS

Let $A \times B$ be the Cartesian product of two sets, A and B.

Definition 1.3.1. A relations ρ from A to B is a subset of $A \times B$, that is, ρ consists of ordered pairs (a, b) such that $a \in A$ and $b \in B$. In particular, if $A = B$, then ρ is said to be a relation in A.

For example, if $A = \{7, 8, 9\}$ and $B = \{7, 8, 9, 10\}$, then $\rho = \{(a, b) | a < b, a \in A, b \in B\}$ is a relation from A to B that consists of the six ordered pairs $(7, 8)$, $(7, 9)$, $(7, 10)$, $(8, 9)$, $(8, 10)$, and $(9, 10)$.

Whenever ρ is a relation and $(x, y) \in \rho$, then x and y are said to be ρ-related. This is denoted by writing $x \rho y$. □

Definition 1.3.2. A relation ρ in a set A is an equivalence relation if the following properties are satisfied:

1. ρ is reflexive, that is, $a \rho a$ for any a in A.
2. ρ is symmetric, that is, if $a \rho b$, then $b \rho a$ for any a, b in A.
3. ρ is transitive, that is, if $a \rho b$ and $b \rho c$, then $a \rho c$ for any a, b, c in A.

If ρ is an equivalence relation in a set A, then for a given a_0 in A, the set

$$C(a_0) = \{a \in A | a_0 \ \rho \ a\},$$

which consists of all elements of A that are ρ-related to a_0, is called an equivalence class of a_0. □

RESULT 1.3.1. $a \in C(a)$ for any a in A. Thus each element of A is an element of an equivalence class.

RESULT 1.3.2. If $C(a_1)$ and $C(a_2)$ are two equivalence classes, then either $C(a_1) = C(a_2)$, or $C(a_1)$ and $C(a_2)$ are disjoint subsets.

It follows from Results 1.3.1 and 1.3.2 that if A is a nonempty set, the collection of distinct ρ-equivalence classes of A forms a partition of A.

As an example of an equivalence relation, consider that $a \rho b$ if and only if a and b are integers such that $a - b$ is divisible by a nonzero integer n. This is the relation of congruence modulo n in the set of integers and is written symbolically as $a \equiv b \pmod{n}$. Clearly, $a \equiv a \pmod{n}$, since $a - a = 0$ is divisible by n. Also, if $a \equiv b \pmod{n}$, then $b \equiv a \pmod{n}$, since if $a - b$ is divisible by n, then so is $b - a$. Furthermore, if $a \equiv b \pmod{n}$ and $b \equiv c \pmod{n}$, then $a \equiv c \pmod{n}$. This is true because if $a - b$ and $b - c$ are both divisible by n, then so is $(a - b) + (b - c) = a - c$. Now, if a_0 is a given integer, then a ρ-equivalence class of a_0 consists of all integers that can be written as $a = a_0 + kn$, where k is an integer. This in this example $C(a_0)$ is the set $\{a_0 + kn | k \in J\}$, where J denotes the set of all integers.

Definition 1.3.3. Let ρ be a relation from A to B. Suppose that ρ has the property that for all x in A, if $x \rho y$ and $x \rho z$, where y and z are elements in B, then $y = z$. Such a relation is called a function. □

Thus a function is a relation ρ such that any two elements in B that are ρ-related to the same x in A must be identical. In other words, to each element x in A, there corresponds only one element y in B. We call y the value of the function at x and denote it by writing $y = f(x)$. The set A is called the domain of the function f, and the set of all values of $f(x)$ for x in A is called the range of f, or the image of A under f, and is denoted by $f(A)$. In this case, we say that f is a function, or a mapping, from A into B. We express this fact by writing $f: A \to B$. Note that $f(A)$ is a subset of B. In particular, if $B = f(A)$, then f is said to be a function from A onto B. In this case, every element b in B has a corresponding element a in A such that $b = f(a)$.

Definition 1.3.4. A function f defined on a set A is said to be a one-to-one function if whenever $f(x_1) = f(x_2)$ for x_1, x_2 in A, one has $x_1 = x_2$. Equivalently, f is a one-to-one function if whenever $x_1 \neq x_2$, one has $f(x_1) \neq f(x_2)$. □

Thus a function $f: A \to B$ is one-to-one if to each y in $f(A)$, there corresponds only one element x in A such that $y = f(x)$. In particular, if f is a one-to-one and onto function, then it is said to provide a one-to-one correspondence between A and B. In this case, the sets A and B are said to be equivalent. This fact is denoted by writing $A \sim B$.

Note that whenever $A \sim B$, there is a function $g: B \to A$ such that if $y = f(x)$, then $x = g(y)$. The function g is called the inverse function of f and

is denoted by f^{-1}. It is easy to see that $A \sim B$ defines an equivalence relation. Properties 1 and 2 in Definition 1.3.2 are obviously true here. As for property 3, if A, B, and C are sets such that $A \sim B$ and $B \sim C$, then $A \sim C$. To show this, let f: $A \to B$ and h: $B \to C$ be one-to-one and onto functions. Then, the composite function $h \circ f$, where $h \circ f(x) = h[f(x)]$, defines a one-to-one correspondence between A and C.

EXAMPLE 1.3.1. The relation $a \rho b$, where a and b are real numbers such that $a = b^2$, is not a function. This is true because both pairs (a, b) and $(a, -b)$ belong to ρ.

EXAMPLE 1.3.2. The relation $a \rho b$, where a and b are real numbers such that $b = 2a^2 + 1$, is a function, since for each a, there is only one b that is ρ-related to a.

EXAMPLE 1.3.3. Let $A = \{x| -1 \le x \le 1\}$, $B = \{x|0 \le x \le 2\}$. Define f: $A \to B$ such that $f(x) = x^2$. Here, f is a function, but is not one-to-one because $f(1) = f(-1) = 1$. Also, f does not map A onto B, since $y = 2$ has no corresponding x in A such that $x^2 = 2$.

EXAMPLE 1.3.4. Consider the relation $x \rho y$, where $y = \arcsin x$, $-1 \le x \le 1$. Here, y is an angle measured in radians whose sine is x. Since there are infinitely many angles with the same sine, ρ is not a function. However, if we restrict the range of y to the set $B = \{y| -\pi/2 \le y \le \pi/2\}$, then ρ becomes a function, which is also one-to-one and onto. This function is the inverse of the sine function $x = \sin y$. We refer to the values of y that belong to the set B as the principal values of $\arcsin x$, which we denote by writing $y = \text{Arcsin } x$. Note that other functions could have also been defined from the arcsine relation. For example, if $\pi/2 \le y \le 3\pi/2$, then $x = \sin y = -\sin z$, where $z = y - \pi$. Since $-\pi/2 \le z \le \pi/2$, then $z = -\text{Arcsin } x$. Thus $y = \pi - \text{Arcsin } x$ maps the set $A = \{x| -1 \le x \le 1\}$ in a one-to-one manner onto the set $C = \{y| \pi/2 \le y \le 3\pi/2\}$.

1.4. FINITE, COUNTABLE, AND UNCOUNTABLE SETS

Let $J_n = \{1, 2, \ldots, n\}$ be a set consisting of the first n positive integers, and let J^+ denote the set of all positive integers.

Definition 1.4.1. A set A is said to be:

1. Finite if $A \sim J_n$ for some positive integer n.
2. Countable if $A \sim J^+$. In this case, the set J^+, or any other set equivalent to it, can be used as an index set for A, that is, the elements of A are assigned distinct indices (subscripts) that belong to J^+. Hence, A can be represented as $A = \{a_1, a_2, \ldots, a_n, \ldots\}$.

3. Uncountable if A is neither finite nor countable. In this case, the elements of A cannot be indexed by J_n for any n, or by J^+. □

EXAMPLE 1.4.1. Let $A = \{1, 4, 9, \ldots, n^2, \ldots\}$. This set is countable, since the function $f: J^+ \to A$ defined by $f(n) = n^2$ is one-to-one and onto. Hence, $A \sim J^+$.

EXAMPLE 1.4.2. Let $A = J$ be the set of all integers. Then A is countable. To show this, consider the function $f: J^+ \to A$ defined by

$$f(n) = \begin{cases} (n+1)/2, & n \text{ odd}, \\ (2-n)/2, & n \text{ even}. \end{cases}$$

It can be verified that f is one-to-one and onto. Hence, $A \sim J^+$.

EXAMPLE 1.4.3. Let $A = \{x | 0 \le x \le 1\}$. This set is uncountable. To show this, suppose that there exists a one-to-one correspondence between J^+ and A. We can then write $A = \{a_1, a_2, \ldots, a_n, \ldots\}$. Let the digit in the nth decimal place of a_n be denoted by b_n $(n = 1, 2, \ldots)$. Define a number c as $c = 0 \cdot c_1 c_2 \cdots c_n \cdots$ such that for each n, $c_n = 1$ if $b_n \neq 1$ and $c_n = 2$ if $b_n = 1$. Now, c belongs to A, since $0 \le c \le 1$. However, by construction, c is different from every a_i in at least one decimal digit $(i = 1, 2, \ldots)$ and hence $c \notin A$, which is a contradiction. Therefore, A is not countable. Since A is not finite either, then it must be uncountable.

This result implies that any subset of R, the set of real numbers, that contains A, or is equivalent to it, must be uncountable. In particular, R is uncountable.

Theorem 1.4.1. Every infinite subset of a countable set is countable.

Proof. Let A be a countable set, and B be an infinite subset of A. Then $A = \{a_1, a_2, \ldots, a_n, \ldots\}$, where the a_i's are distinct elements. Let n_1 be the smallest positive integer such that $a_{n_1} \in B$. Let $n_2 > n_1$ be the next smallest integer such that $a_{n_2} \in B$. In general, if $n_1 < n_2 < \cdots < n_{k-1}$ have been chosen, let n_k be the smallest integer greater than n_{k-1} such that $a_{n_k} \in B$. Define the function $f: J^+ \to B$ such that $f(k) = a_{n_k}$, $k = 1, 2, \ldots$. This function is one-to-one and onto. Hence, B is countable. □

Theorem 1.4.2. The union of two countable sets is countable.

Proof. Let A and B be countable sets. Then they can be represented as $A = \{a_1, a_2, \ldots, a_n, \ldots\}$, $B = \{b_1, b_2, \ldots, b_n, \ldots\}$. Define $C = A \cup B$. Consider the following two cases:

i. A and B are disjoint.

ii. A and B are not disjoint.

In case i, let us write C as $C = \{a_1, b_1, a_2, b_2, \ldots, a_n, b_n, \ldots\}$. Consider the function f: $J^+ \to C$ such that

$$f(n) = \begin{cases} a_{(n+1)/2}, & n \text{ odd}, \\ b_{n/2}, & n \text{ even}. \end{cases}$$

It can be verified that f is one-to-one and onto. Hence, C is countable.

Let us now consider case ii. If $A \cap B \neq \varnothing$, then some elements of C, namely those in $A \cap B$, will appear twice. Hence, there exists a set $E \subset J^+$ such that $E \sim C$. Thus C is either finite or countable. Since $C \supset A$ and A is infinite, C must be countable. □

Corollary 1.4.1. If $A_1, A_2, \ldots, A_n, \ldots,$ are countable sets, then $\bigcup_{i=1}^{\infty} A_i$ is countable.

Proof. The proof is left as an exercise. □

Theorem 1.4.3. Let A and B be two countable sets. Then their Cartesian product $A \times B$ is countable.

Proof. Let us write A as $A = (a_1, a_2, \ldots, a_n, \ldots\}$. For a given $a \in A$, define (a, B) as the set

$$(a, B) = \{(a, b) | b \in B\}.$$

Then $(a, B) \sim B$ and hence (a, B) is countable.

However,

$$A \times B = \bigcup_{i=1}^{\infty} (a_i, B).$$

Thus by Corollary 1.4.1, $A \times B$ is countable. □

Corollary 1.4.2. If $A_1, A_2, \ldots, A_n$ are countable sets, then their Cartesian product $\times_{i=1}^{n} A_i$ is countable.

Proof. The proof is left as an exercise. □

Corollary 1.4.3. The set Q of all rational numbers is countable.

Proof. By definition, a rational number is a number of the form m/n, where m and n are integers with $n \neq 0$. Thus $Q \sim \tilde{Q}$, where

$$\tilde{Q} = \{(m, n) | m, n \text{ are integers and } n \neq 0\}.$$

Since $\tilde{Q}$ is an infinite subset of $J \times J$, where J is the set of all integers, which is countable as was seen in Example 1.4.2, then by Theorems 1.4.1 and 1.4.3, $\tilde{Q}$ is countable and so is Q. □

REMARK 1.4.1. Any real number that cannot be expressed as a rational number is called an irrational number. For example, $\sqrt{2}$ is an irrational number. To show this, suppose that there exist integers, m and n, such that $\sqrt{2} = m/n$. We may consider that m/n is written in its lowest terms, that is, m and n have no common factors other than unity. In particular, m and n, cannot both be even. Now, $m^2 = 2n^2$. This implies that m^2 is even. Hence, m is even and can therefore be written as $m = 2m'$. It follows that $n^2 = m^2/2 = 2m'^2$. Consequently, n^2, and hence n, is even. This contradicts the fact that m and n are not both even. Thus $\sqrt{2}$ must be an irrational number.

1.5. BOUNDED SETS

Let us consider the set R of real numbers.

Definition 1.5.1. A set $A \subset R$ is said to be:

1. Bounded from above if there exists a number q such that $x \leq q$ for all x in A. This number is called an upper bound of A.
2. Bounded from below if there exists a number p such that $x \geq p$ for all x in A. The number p is called a lower bound of A.
3. Bounded if A has an upper bound q and a lower bound p. In this case, there exists a nonnegative number r such that $-r \leq x \leq r$ for all x in A. This number is equal to $\max(|p|, |q|)$. □

Definition 1.5.2. Let $A \subset R$ be a set bounded from above. If there exists a number l that is an upper bound of A and is less than or equal to any other upper bound of A, then l is called the least upper bound of A and is denoted by $\text{lub}(A)$. Another name for $\text{lub}(A)$ is the supremum of A and is denoted by $\sup(A)$. □

Definition 1.5.3. Let $A \subset R$ be a set bounded from below. If there exists a number g that is a lower bound of A and is greater than or equal to any other lower bound of A, then g is called the greatest lower bound and is denoted by $\text{glb}(A)$. The infimum of A, denoted by $\inf(A)$, is another name for $\text{glb}(A)$. □

The least upper bound of A, if it exists, is unique, but it may or may not belong to A. The same is true for $\text{glb}(A)$. The proof of the following theorem is omitted and can be found in Rudin (1964, Theorem 1.36).

Theorem 1.5.1. Let $A \subset R$ be a nonempty set.

1. If A is bounded from above, then $\text{lub}(A)$ exists.
2. If A is bounded from below, then $\text{glb}(A)$ exists.

EXAMPLE 1.5.1. Let $A = \{x | x < 0\}$. Then $\text{lub}(A) = 0$, which does not belong to A.

EXAMPLE 1.5.2. Let $A = \{1/n | n = 1, 2, \ldots\}$. Then $\text{lub}(A) = 1$ and $\text{glb}(A) = 0$. In this case, $\text{lub}(A)$ belongs to A, but $\text{glb}(A)$ does not.

1.6. SOME BASIC TOPOLOGICAL CONCEPTS

The field of topology is an abstract study that evolved as an independent discipline in response to certain problems in classical analysis and geometry. It provides a unifying theory that can be used in many diverse branches of mathematics. In this section, we present a brief account of some basic definitions and results in the so-called *point-set topology*.

Definition 1.6.1. Let A be a set, and let $\mathscr{F} = \{B_\alpha\}$ be a family of subsets of A. Then $\mathscr{F}$ is a *topology* in A if it satisfies the following properties:

1. The union of any number of members of $\mathscr{F}$ is also a member of $\mathscr{F}$.
2. The intersection of a finite number of members of $\mathscr{F}$ is also a member of $\mathscr{F}$.
3. Both A and the empty set $\varnothing$ are members of $\mathscr{F}$. □

Definition 1.6.2. Let $\mathscr{F}$ be a topology in a set A. Then the pair $(A, \mathscr{F})$ is called a *topological space*. □

Definition 1.6.3. Let $(A, \mathscr{F})$ be a topological space. Then the members of $\mathscr{F}$ are called the *open sets* of the topology $\mathscr{F}$. □

Definition 1.6.4. Let $(A, \mathscr{F})$ be a topological space. A neighborhood of a point $p \in A$ is any open set (that is, a member of $\mathscr{F}$) that contains p. In particular, if $A = R$, the set of real numbers, then a neighborhood of $p \in R$ is an open set of the form $N_r(p) = \{q | \ |q - p| < r\}$ for some $r > 0$. □

Definition 1.6.5. Let $(A, \mathscr{F})$ be a topological space. A family $G = \{B_\alpha\} \subset \mathscr{F}$ is called a *basis* for $\mathscr{F}$ if each open set (that is, member of $\mathscr{F}$) is the union of members of G. □

On the basis of this definition, it is easy to prove the following theorem.

Theorem 1.6.1. Let $(A, \mathscr{F})$ be a topological space, and let G be a basis for $\mathscr{F}$. Then a set $B \subset A$ is open (that is, a member of $\mathscr{F}$) if and only if for each $p \in B$, there is a $U \in G$ such that $p \in U \subset B$.

For example, if $A = R$, then $G = \{N_r(p) | p \in R,\ r > 0\}$ is a basis for the topology in R. It follows that a set $B \subset R$ is open if for every point p in B, there exists a neighborhood $N_r(p)$ such that $N_r(p) \subset B$.

Definition 1.6.6. Let $(A, \mathscr{F})$ be a topological space. A set $B \subset A$ is closed if $\overline{B}$, the complement of B with respect to A, is an open set. □

It is easy to show that closed sets of a topological space $(A, \mathscr{F})$ satisfy the following properties:

1. The intersection of any number of closed sets is closed.
2. The union of a finite number of closed sets is closed.
3. Both A and the empty set $\varnothing$ are closed.

Definition 1.6.7. Let $(A, \mathscr{F})$ be a topological space. A point $p \in A$ is said to be a limit point of a set $B \subset A$ if every neighborhood of p contains at least one element of B distinct from p. Thus, if $U(p)$ is any neighborhood of p, then $U(p) \cap B$ is a nonempty set that contains at least one element besides p. In particular, if $A = R$, the set of real numbers, then p is a limit point of a set $B \subset R$ if for any $r > 0$, $N_r(p) \cap [B - \{p\}] \neq \varnothing$, where $\{p\}$ denotes a set consisting of just p. □

Theorem 1.6.2. Let p be a limit point of a set $B \subset R$. Then every neighborhood of p contains infinitely many points of B.

Proof. The proof is left to the reader. □

The next theorem is a fundamental theorem in set theory. It is originally due to Bernhard Bolzano (1781–1848), though its importance was first recognized by Karl Weierstrass (1815–1897). The proof is omitted and can be found, for example, in Zaring (1967, Theorem 4.62).

Theorem 1.6.3 (Bolzano–Weierstrass). Every bounded infinite subset of R, the set of real numbers, has at least one limit point.

Note that a limit point of a set B may not belong to B. For example, the set $B = \{1/n | n = 1, 2, \ldots\}$ has a limit point equal to zero, which does not belong to B. It can be seen here that any neighborhood of 0 contains infinitely many points of B. In particular, if r is a given positive number, then all elements of B of the form $1/n$, where $n > 1/r$, belong to $N_r(0)$. From Theorem 1.6.2 it can also be concluded that a finite set cannot have limit points.

Limit points can be used to describe closed sets, as can be seen from the following theorem.

Theorem 1.6.4. A set B is closed if and only if every limit point of B belongs to B.

Proof. Suppose that B is closed. Let p be a limit point of B. If $p \notin B$, then $p \in \overline{B}$, which is open. Hence, there exists a neighborhood $U(p)$ of p contained inside $\overline{B}$ by Theorem 1.6.1. This means that $U(p) \cap B = \emptyset$, a contradiction, since p is a limit point of B (see Definition 1.6.7). Therefore, p must belong to B. Vice versa, if every limit point of B is in B, then B must be closed. To show this, let p be any point in $\overline{B}$. Then, p is not a limit point of B. Therefore, there exists a neighborhood $U(p)$ such that $U(p) \subset \overline{B}$. This means that $\overline{B}$ is open and hence B is closed. □

It should be noted that a set does not have to be either open or closed; if it is closed, it does not have to be open, and vice versa. Also, a set may be both open and closed.

EXAMPLE 1.6.1. $B = \{x | 0 < x < 1\}$ is an open subset of R, but is not closed, since both 0 and 1 are limit points of B, but do not belong to it.

EXAMPLE 1.6.2. $B = \{x | 0 \leq x \leq 1\}$ is closed, but is not open, since any neighborhood of 0 or 1 is not contained in B.

EXAMPLE 1.6.3. $B = \{x | 0 < x \leq 1\}$ is not open, because any neighborhood of 1 is not contained in B. It is also not closed, because 0 is a limit point that does not belong to B.

EXAMPLE 1.6.4. The set R is both open and closed.

EXAMPLE 1.6.5. A finite set is closed because it has no limit points, but is obviously not open.

Definition 1.6.8. A subset B of a topological space $(A, \mathscr{F})$ is *disconnected* if there exist open subsets C and D of A such that $B \cap C$ and $B \cap D$ are disjoint nonempty sets whose union is B. A set is *connected* if it is not disconnected. □

The set of all rationals Q is disconnected, since $\{x | x > \sqrt{2}\} \cap Q$ and $\{x | x < \sqrt{2}\} \cap Q$ are disjoint nonempty sets whose union is Q. On the other hand, all intervals in R (open, closed, or half-open) are connected.

Definition 1.6.9. A collection of sets $\{B_\alpha\}$ is said to be a *covering* of a set A if the union $\bigcup_\alpha B_\alpha$ contains A. If each B_α is an open set, then $\{B_\alpha\}$ is called an *open covering*.

Definition 1.6.10. A set A in a topological space is *compact* if each open covering $\{B_\alpha\}$ of A has a finite subcovering, that is, there is a finite subcollection $B_{\alpha_1}, B_{\alpha_2}, \ldots, B_{\alpha_n}$ of $\{B_\alpha\}$ such that $A \subset \bigcup_{i=1}^{n} B_{\alpha_i}$. □

The concept of compactness is motivated by the classical *Heine–Borel theorem*, which characterizes compact sets in R, the set of real numbers, as closed and bounded sets.

Theorem 1.6.5 (Heine–Borel). A set $B \subset R$ is compact if and only if it is closed and bounded.

Proof. See, for example, Zaring (1967, Theorem 4.78). □

Thus, according to the Heine–Borel theorem, every closed and bounded interval $[a, b]$ is compact.

1.7. EXAMPLES IN PROBABILITY AND STATISTICS

EXAMPLE 1.7.1. In probability theory, events are considered as subsets in a sample space Ω, which consists of all the possible outcomes of an experiment. A Borel field of events (also called a σ-field) in Ω is a collection $\mathscr{B}$ of events with the following properties:

- **i.** $\Omega \in \mathscr{B}$.
- **ii.** If $E \in \mathscr{B}$, then $\overline{E} \in \mathscr{B}$, where $\overline{E}$ is the complement of E.
- **iii.** If $E_1, E_2, \ldots, E_n, \ldots$ is a countable collection of events in $\mathscr{B}$, then $\bigcup_{i=1}^{\infty} E_i$ belongs to $\mathscr{B}$.

The probability of an event E is a number denoted by $P(E)$ that has the following properties:

- **i.** $0 \le P(E) \le 1$.
- **ii.** $P(\Omega) = 1$.
- **iii.** If $E_1, E_2, \ldots, E_n, \ldots$ is a countable collection of disjoint events in $\mathscr{B}$, then

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i).$$

By definition, the triple $(\Omega, \mathscr{B}, P)$ is called a probability space.

EXAMPLE 1.7.2 . A random variable X defined on a probability space $(\Omega, \mathscr{B}, P)$ is a function $X\colon \Omega \to A$, where A is a nonempty set of real numbers. For any real number x, the set $E = \{\omega \in \Omega | X(\omega) \le x\}$ is an

element of $\mathscr{B}$. The probability of the event E is called the cumulative distribution function of X and is denoted by $F(x)$. In statistics, it is customary to write just X instead of $X(\omega)$. We thus have

$$F(x) = P(X \leq x).$$

This concept can be extended to several random variables: Let $X_1, X_2, \ldots, X_n$ be n random variables. Define the event $A_i = \{\omega \in \Omega | X_i(\omega) \leq x_i\}$, $i = 1, 2, \ldots, n$. Then, $P(\bigcap_{i=1}^{n} A_i)$, which can be expressed as

$$F(x_1, x_2, \ldots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n),$$

is called the joint cumulative distribution function of $X_1, X_2, \ldots, X_n$. In this case, the n-tuple $(X_1, X_2, \ldots, X_n)$ is said to have a multivariate distribution.

A random variable X is said to be discrete, or to have a discrete distribution, if its range is finite or countable. For example, the binomial random variable is discrete. It represents the number of successes in a sequence of n independent trials, in each of which there are two possible outcomes: success or failure. The probability of success, denoted by p_n, is the same in all the trials. Such a sequence of trials is called a Bernoulli sequence. Thus the possible values of this random variable are $0, 1, \ldots, n$.

Another example of a discrete random variable is the Poisson, whose possible values are $0, 1, 2, \ldots$. It is considered to be the limit of a binomial random variable as $n \to \infty$ in such a way that $np_n \to \lambda > 0$. Other examples of discrete random variables include the discrete uniform, geometric, hypergeometric, and negative binomial (see, for example, Fisz, 1963; Johnson and Kotz, 1969; Lindgren 1976; Lloyd, 1980).

A random variable X is said to be continuous, or to have a continuous distribution, if its range is an uncountable set, for example, an interval. In this case, the cumulative distribution function $F(x)$ of X is a continuous function of x on the set R of all real numbers. If, in addition, $F(x)$ is differentiable, then its derivative is called the density function of X. One of the best-known continuous distributions is the normal. A number of continuous distributions are derived in connection with it, for example, the chi-squared, F, Rayleigh, and t distributions. Other well-known continuous distributions include the beta, continuous uniform, exponential, and gamma distributions (see, for example, Fisz, 1963; Johnson and Kotz, 1970a, b).

EXAMPLE 1.7.3. Let $f(x, \theta)$ denote the density function of a continuous random variable X, where θ represents a set of unknown parameters that identify the distribution of X. The range of X, which consists of all possible values of X, is referred to as a population and denoted by P_X. Any subset of n elements from P_X forms a sample of size n. This sample is actually an element in the Cartesian product P_X^n. Any real-valued function defined on P_X^n is called a statistic. We denote such a function by $g(X_1, X_2, \ldots, X_n)$, where each X_i has the same distribution as X. Note that this function is a random variable whose values do not depend on θ. For example, the sample mean $\bar{X} = \sum_{i=1}^{n} X_i/n$ and the sample variance $S^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2/(n-1)$

are statistics. We adopt the convention that whenever a particular sample of size n is chosen (or observed) from P_X, the elements in that sample are written using lowercase letters, for example, $x_1, x_2, \ldots, x_n$. The corresponding value of a statistic is written as $g(x_1, x_2, \ldots, x_n)$.

EXAMPLE 1.7.4. Two random variables, X and Y, are said to be equal in distribution if they have the same cumulative distribution function. This fact is denoted by writing $X \stackrel{d}{=} Y$. The same definition applies to random variables with multivariate distributions. We note that $\stackrel{d}{=}$ is an equivalence relation, since it satisfies properties 1, 2, and 3 in Definition 1.3.2. The first two properties are obviously true. As for property 3, if $X \stackrel{d}{=} Y$ and $Y \stackrel{d}{=} Z$, then $X \stackrel{d}{=} Z$, which implies that all three random variables have the same cumulative distribution function. This equivalence relation is useful in nonparametric statistics (see Randles and Wolfe, 1979). For example, it can be shown that if X has a distribution that is symmetric about some number μ, then $X - \mu \stackrel{d}{=} \mu - X$. Also, if $X_1, X_2, \ldots, X_n$ are independent and identically distributed random variables, and if $(m_1, m_2, \ldots, m_n)$ is any permutation of the n-tuple $(1, 2, \ldots, n)$, then $(X_1, X_2, \ldots, X_n) \stackrel{d}{=} (X_{m_1}, X_{m_2}, \ldots, X_{m_n})$. In this case, we say that the collection of random variables $X_1, X_2, \ldots, X_n$ is exchangeable.

EXAMPLE 1.7.5. Consider the problem of testing the null hypothesis H_0: $\theta \leq \theta_0$ versus the alternative hypothesis H_a: $\theta > \theta_0$, where θ is some unknown parameter that belongs to a set A. Let T be a statistic used in making a decision as to whether H_0 should be rejected or not. This statistic is appropriately called a test statistic.

Suppose that H_0 is rejected if $T > t$, where t is some real number. Since the distribution of T depends on θ, then the probability $P(T > t)$ is a function of θ, which we denote by $\pi(\theta)$. Thus π: $A \to [0,1]$. Let B_0 be a subset of A defined as $B_0 = \{\theta \in A \mid \theta \leq \theta_0\}$. By definition, the size of the test is the least upper bound of the set $\pi(B_0)$. This probability is denoted by α and is also called the level of significance of the test. We thus have

$$\alpha = \sup_{\theta \leq \theta_0} \pi(\theta).$$

To learn more about the above examples and others, the interested reader may consider consulting some of the references listed in the annotated bibliography.

FURTHER READING AND ANNOTATED BIBLIOGRAPHY

Bronshtein, I. N., and K. A. Semendyayev (1985). *Handbook of Mathematics* (English translation edited by K. A. Hirsch). Van Nostrand Reinhold, New York. (Section 4.1 in this book gives basic concepts of set theory; Chap. 5 provides a brief introduction to probability and mathematical statistics.)

Dugundji, J. (1966). *Topology*. Allyn and Bacon, Boston. (Chap. 1 deals with elementary set theory; Chap. 3 presents some basic topological concepts that complements the material given in Section 1.6.)

Fisz, M. (1963). *Probability Theory and Mathematical Statistics*, 3rd ed. Wiley, New York. (Chap. 1 discusses random events and axioms of the theory of probability; Chap. 2 introduces the concept of a random variable; Chap. 5 investigates some probability distributions.)

Hardy, G. H. (1955). *A Course of Pure Mathematics*, 10th ed. The University Press, Cambridge, England. (Chap. 1 in this classic book is recommended reading for understanding the real number system.)

Harris, B. (1966). *Theory of Probability*. Addison-Wesley, Reading, Massachusetts. (Chaps. 2 and 3 discuss some elementary concepts in probability theory as well as in distribution theory. Many exercises are provided.)

Hogg, R. V., and A. T. Craig (1965). *Introduction to Mathematical Statistics*, 2nd ed. Macmillan, New York. (Chap. 1 is an introduction to distribution theory; examples of some special distributions are given in Chap. 3; Chap. 10 considers some aspects of hypothesis testing that pertain to Example 1.7.5.)

Johnson, N. L., and S. Kotz (1969). *Discrete Distributions*. Houghton Mifflin, Boston. (This is the first volume in a series of books on statistical distributions. It is an excellent source for getting detailed accounts of the properties and uses of these distributions. This volume deals with discrete distributions, including the binomial in Chap. 3, the Poisson in Chap. 4, the negative binomial in Chap. 5, and the hypergeometric in Chap. 6.)

Johnson, N. L., and S. Kotz (1970a). *Continuous Univariate Distributions—1*. Houghton Mifflin, Boston. (This volume covers continuous distributions, including the normal in Chap. 13, lognormal in Chap. 14, Cauchy in Chap. 16, gamma in Chap. 17, and the exponential in Chap. 18.)

Johnson, N. L., and S. Kotz (1970b). *Continuous Univariate Distributions—2*. Houghton Mifflin, Boston. (This is a continuation of Vol. 2 on continuous distributions. Chaps. 24, 25, 26, and 27 discuss the beta, continuous uniforms, F, and t distributions, respectively.)

Johnson, P. E. (1972). *A History of Set Theory*. Prindle, Weber, and Schmidt, Boston. (This book presents a historical account of set theory as was developed by Georg Cantor.)

Lindgren, B. W. (1976). *Statistical Theory*, 3rd ed. Macmillan, New York. (Sections 1.1, 1.2, 2.1, 3.1, 3.2, and 3.3 present introductory material on probability models and distributions; Chap. 6 discusses test of hypothesis and statistical inference.)

Lloyd, E. (1980). *Handbook of Applicable Mathematics*, Vol. II. Wiley, New York. (This is the second volume in a series of six volumes designed as texts of mathematics for professionals. Chaps. 1, 2, and 3 present expository material on probability; Chaps. 4 and 5 discuss random variables and their distributions.)

Randles, R. H., and D. A. Wolfe (1979). *Introduction to the Theory of Nonparametric Statistics*. Wiley, New York. (Section 1.3 in this book discusses the "equal in distribution" property mentioned in Example 1.7.4.)

Rudin, W. (1964). *Principles of Mathematical Analysis*, 2nd ed. McGraw-Hill, New York. (Chap. 1 discusses the real number system; Chap. 2 deals with countable, uncountable, and bounded sets and pertains to Sections 1.4, 1.5, and 1.6.)

Stoll, R. R. (1963). *Set Theory and Logic*. W. H. Freeman, San Francisco. (Chap. 1 is an introduction to set theory; Chap. 2 discusses countable sets; Chap. 3 is useful in understanding the real number system.)

Tucker, H. G. (1962). *Probability and Mathematical Statistics*. Academic Press, New York. (Chaps. 1, 3, 4, and 6 discuss basic concepts in elementary probability and distribution theory.)

Vilenkin, N. Y. (1968). *Stories about Sets*. Academic Press, New York. (This is an interesting book that presents various notions of set theory in an informal and delightful way. It contains many unusual stories and examples that make the learning of set theory rather enjoyable.)

Zaring, W. M. (1967). *An Introduction to Analysis*. Macmillan, New York. (Chap. 2 gives an introduction to set theory; Chap. 3 discusses functions and relations.)

EXERCISES

In Mathematics

1.1. Verify Results 1.2.3–1.2.12.

1.2. Verify Results 1.2.13–1.2.16.

1.3. Let A, B, and C be sets such that $A \cap B \subset \overline{C}$ and $A \cup C \subset B$. Show that A and C are disjoint.

1.4. Let A, B, and C be sets such that $C = (A - B) \cup (B - A)$. The set C is called the symmetric difference of A and B and is denoted by $A \triangle B$. Show that

(a) $A \triangle B = A \cup B - A \cap B$

(b) $A \triangle (B \triangle D) = (A \triangle B) \triangle D$, where D is any set.

(c) $A \cap (B \triangle D) = (A \cap B) \triangle (A \cap D)$, where D is any set.

1.5. Let $A = J^+ \times J^+$, where J^+ is the set of positive integers. Define a relation ρ in A as follows: If (m_1, n_1) and (m_2, n_2) are elements in A, then $(m_1, n_1)\,\rho\,(m_2, n_2)$ if $m_1 n_2 = n_1 m_2$. Show that ρ is an equivalence relation and describe its equivalence classes.

1.6. Let A be the same set as in Exercise 1.5. Show that the following relation is an equivalence relation: $(m_1, n_1)\,\rho\,(m_2, n_2)$ if $m_1 + n_2 = n_1 + m_2$. Draw the equivalence class of $(1, 2)$.

1.7. Consider the set $A = \{(-2, -5), (-1, -3), (1, 2), (3, 10)\}$. Show that A defines a function.

1.8. Let A and B be two sets and f be a function defined on A such that $f(A) \subset B$. If $A_1, A_2, \ldots, A_n$ are subsets of A, then show that:

(a) $f(\bigcup_{i=1}^{n} A_i) = \bigcup_{i=1}^{n} f(A_i)$.

(b) $f(\bigcap_{i=1}^{n} A_i) \subset \bigcap_{i=1}^{n} f(A_i)$.
Under what conditions are the two sides in (b) equal?

1.9. Prove Corollary 1.4.1.

1.10. Prove Corollary 1.4.2.

1.11. Show that the set $A = \{3, 9, 19, 33, 51, 73, \ldots\}$ is countable.

1.12. Show that $\sqrt{3}$ is an irrational number.

1.13. Let a, b, c, and d be rational numbers such that $a + \sqrt{b} = c + \sqrt{d}$. Then, either
(a) $a = c, b = d$, or
(b) b and d are both squares of rational numbers.

1.14. Let $A \subset R$ be a nonempty set bounded from below. Define $-A$ to be the set $\{-x | x \in A\}$. Show that $\inf(A) = -\sup(-A)$.

1.15. Let $A \subset R$ be a closed and bounded set, and let $\sup(A) = b$. Show that $b \in A$.

1.16. Prove Theorem 1.6.2.

1.17. Let $(A, \mathscr{F})$ be a topological space. Show that $G \subset \mathscr{F}$ is a basis for $\mathscr{F}$ in and only if for each $B \in \mathscr{F}$ and each $p \in B$, there is a $U \in G$ such that $p \in U \subset B$.

1.18. Show that if A and B are closed sets, then $A \cup B$ is a closed set.

1.19. Let $B \subset A$ be a closed subset of a compact set A. Show that B is compact.

1.20. Is a compact subset of a compact set necessarily closed?

In Statistics

1.21. Let X be a random variable. Consider the following events:

$$A_n = \{\omega \in \Omega | X(\omega) < x + 3^{-n}\}, \qquad n = 1, 2, \ldots,$$
$$B_n = \{\omega \in \Omega | X(\omega) \leq x - 3^{-n}\}, \qquad n = 1, 2, \ldots,$$
$$A = \{\omega \in \Omega | X(\omega) \leq x\},$$
$$B = \{\omega \in \Omega | X(\omega) < x\},$$

where x is a real number. Show that for any x,

(a) $\bigcap_{n=1}^{\infty} A_n = A$;

(b) $\bigcup_{n=1}^{\infty} B_n = B$.

1.22. Let X be a nonnegative random variable such that $E(X) = \mu$ is finite, where $E(X)$ denotes the expected value of X. The following inequality, known as *Markov's inequality*, is true:

$$P(X \geq h) \leq \frac{\mu}{h},$$

where h is any positive number. Consider now a Poisson random variable with parameter λ.

(a) Find an upper bound on the probability $P(X \geq 2)$ using Markov's inequality.

(b) Obtain the exact probability value in (a), and demonstrate that it is smaller than the corresponding upper bound in Markov's inequality.

1.23. Let X be a random variable whose expected value μ and variance σ^2 exist. Show that for any positive constants c and k,

(a) $P(|X - \mu| \geq c) \leq \sigma^2/c^2$,

(b) $P(|X - \mu| \geq k\sigma) \leq 1/k^2$,

(c) $P(|X - \mu| < k\sigma) \geq 1 - 1/k^2$.

The preceding three inequalities are equivalent versions of the so-called *Chebyshev's inequality*.

1.24. Let X be a continuous random variable with the density function

$$f(x) = \begin{cases} 1 - |x|, & -1 < x < 1, \\ 0 & \text{elsewhere}. \end{cases}$$

By definition, the density function of X is a nonnegative function such that $F(x) = \int_{-\infty}^{x} f(t)\,dt$, where $F(x)$ is the cumulative distribution function of X.

(a) Apply Markov's inequality to finding upper bounds on the following probabilities: (i) $P(|X| \geq \frac{1}{2})$; (ii) $P(|X| > \frac{1}{3})$.

(b) Compute the exact value of $P(|X| \geq \frac{1}{2})$, and compare it against the upper bound in (a)(i).

1.25. Let $X_1, X_2, \ldots, X_n$ be n continuous random variables. Define the random variables $X_{(1)}$ and $X_{(n)}$ as

$$X_{(1)} = \min_{1 \leq i \leq n} \{X_1, X_2, \ldots, X_n\},$$

$$X_{(n)} = \max_{1 \leq i \leq n} \{X_1, X_2, \ldots, X_n\}.$$

Show that for any x,

(a) $P(X_{(1)} \geq x) = P(X_1 \geq x, X_2 \geq x, \ldots, X_n \geq x)$,

(b) $P(X_{(n)} \leq x) = P(X_1 \leq x, X_2 \leq x, \ldots, X_n \leq x)$.

In particular, if $X_1, X_2, \ldots, X_n$ form a sample of size n from a population with a cumulative distribution function $F(x)$, show that

(c) $P(X_{(1)} \leq x) = 1 - [1 - F(x)]^n$,

(d) $P(X_{(n)} \leq x) = [F(x)]^n$.

The statistics $X_{(1)}$ and $X_{(n)}$ are called the first-order and nth-order statistics, respectively.

1.26. Suppose that we have a sample of size $n = 5$ from a population with an exponential distribution whose density function is

$$f(x) = \begin{cases} 2e^{-2x}, & x > 0, \\ 0 & \text{elsewhere}. \end{cases}$$

Find the value of $P(2 \leq X_{(1)} \leq 3)$.

CHAPTER 2

Basic Concepts in Linear Algebra

In this chapter we present some fundamental concepts concerning vector spaces and matrix algebra. The purpose of the chapter is to familiarize the reader with these concepts, since they are essential to the understanding of some of the remaining chapters. For this reason, most of the theorems in this chapter will be stated without proofs. There are several excellent books on linear algebra that can be used for a more detailed study of this subject (see the bibliography at the end of this chapter).

In statistics, matrix algebra is used quite extensively, especially in linear models and multivariate analysis. The books by Basilevsky (1983), Graybill (1983), Magnus and Neudecker (1988), and Searle (1982) include many applications of matrices in these areas.

In this chapter, as well as in the remainder of the book, elements of the set of real numbers, R, are sometimes referred to as scalars. The Cartesian product $\times_{i=1}^{n} R$ is denoted by R^n, which is also known as the n-dimensional Euclidean space. Unless otherwise stated, all matrix elements are considered to be real numbers.

2.1. VECTOR SPACES AND SUBSPACES

A vector space over R is a set V of elements called vectors together with two operations, addition and scalar multiplication, that satisfy the following conditions:

1. $\mathbf{u} + \mathbf{v}$ is an element of V for all $\mathbf{u}, \mathbf{v}$ in V.
2. If α is a scalar and $\mathbf{u} \in V$, then $\alpha \mathbf{u} \in V$.
3. $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$ for all $\mathbf{u}, \mathbf{v}$ in V.
4. $\mathbf{u} + (\mathbf{v} + \mathbf{w}) = (\mathbf{u} + \mathbf{v}) + \mathbf{w}$ for all $\mathbf{u}, \mathbf{v}, \mathbf{w}$ in V.
5. There exists an element $\mathbf{0} \in V$ such that $\mathbf{0} + \mathbf{u} = \mathbf{u}$ for all $\mathbf{u}$ in V. This element is called the zero vector.

6. For each $\mathbf{u} \in V$ there exists a $\mathbf{v} \in V$ such that $\mathbf{u} + \mathbf{v} = \mathbf{0}$.
7. $\alpha(\mathbf{u} + \mathbf{v}) = \alpha\mathbf{u} + \alpha\mathbf{v}$ for any scalar α and any $\mathbf{u}$ and $\mathbf{v}$ in V.
8. $(\alpha + \beta)\mathbf{u} = \alpha\mathbf{u} + \beta\mathbf{u}$ for any scalars α and β and any $\mathbf{u}$ in V.
9. $\alpha(\beta\mathbf{u}) = (\alpha\beta)\mathbf{u}$ for any scalars α and β and any $\mathbf{u}$ in V.
10. $1\mathbf{u} = \mathbf{u}$ for any $\mathbf{u} \in V$.

EXAMPLE 2.1.1. A familiar example of a vector space is the n-dimensional Euclidean space R^n. Here, addition and multiplication are defined as follows: If $(u_1, u_2, \ldots, u_n)$ and $(v_1, v_2, \ldots, v_n)$ are two elements in R^n, then their sum is defined as $(u_1 + v_1, u_2 + v_2, \ldots, u_n + v_n)$. If α is a scalar, then $\alpha(u_1, u_2, \ldots, u_n) = (\alpha u_1, \alpha u_2, \ldots, \alpha u_n)$.

EXAMPLE 2.1.2. Let V be the set of all polynomials in x of degree less than or equal to k. Then V is a vector space. Any element in V can be expressed as $\sum_{i=0}^{k} a_i x^i$, where the a_i's are scalars.

EXAMPLE 2.1.3. Let V be the set of all functions defined on the closed interval $[-1, 1]$. Then V is a vector space. It can be seen that $f(x) + g(x)$ and $\alpha f(x)$ belong to V, where $f(x)$ and $g(x)$ are elements in V and α is any scalar.

EXAMPLE 2.1.4. The set V of all nonnegative functions defined on $[-1, 1]$ is not a vector space, since if $f(x) \in V$ and α is a negative scalar, then $\alpha f(x) \notin V$.

EXAMPLE 2.1.5. Let V be the set of all points (x, y) on a straight line given by the equation $2x - y + 1 = 0$. Then V is not a vector space. This is because if (x_1, y_1) and (x_2, y_2) belong to V, then $(x_1 + x_2, y_1 + y_2) \notin V$, since $2(x_1 + x_2) - (y_1 + y_2) + 1 = -1 \neq 0$. Alternatively, we can state that V is not a vector space because the zero element $(0, 0)$ does not belong to V. This violates condition 5 for a vector space.

A subset W of a vector space V is said to form a vector subspace if W itself is a vector space. Equivalently, W is a subspace if whenever $\mathbf{u}, \mathbf{v} \in W$ and α is a scalar, then $\mathbf{u} + \mathbf{v} \in W$ and $\alpha\mathbf{u} \in W$. For example, the set W of all continuous functions defined on $[-1, 1]$ is a vector subspace of V in Example 2.1.3. Also, the set of all points on the straight line $y - 2x = 0$ is a vector subspace of R^2. However, the points on any straight line in R^2 not going through the origin $(0, 0)$ do not form a vector subspace, as was seen in Example 2.1.5.

Definition 2.1.1. Let V be a vector space, and $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$ be a collection of n elements in V. These elements are said to be linearly dependent if there exist n scalars $\alpha_1, \alpha_2, \ldots, \alpha_n$, not all equal to zero, such that $\sum_{i=1}^{n} \alpha_i \mathbf{u}_i = \mathbf{0}$. If, however, $\sum_{i=1}^{n} \alpha_i \mathbf{u}_i = \mathbf{0}$ is true only when all the α_i's are zero, then

$\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$ are linearly independent. It should be noted that if $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$ are linearly independent, then none of them can be zero. If, for example, $\mathbf{u}_1 = \mathbf{0}$, then $\alpha \mathbf{u}_1 + 0\mathbf{u}_2 + \cdots + 0\mathbf{u}_n = \mathbf{0}$ for any $\alpha \neq 0$, which implies that the $\mathbf{u}_i$'s are linearly dependent, a contradiction. □

From the preceding definition we can say that a collection of n elements in a vector space are linearly dependent if at least one element in this collection can be expressed as a linear combination of the remaining $n-1$ elements. If no element, however, can be expressed in this fashion, then the n elements are linearly independent. For example, in $R^3, (1, 2, -2), (-1, 0, 3)$, and $(1, 4, -1)$ are linearly dependent, since $2(1, 2, -2) + (-1, 0, 3) - (1, 4, -1) = \mathbf{0}$. On the other hand, it can be verified that $(1, 1, 0)$, $(1, 0, 2)$, and $(0, 1, 3)$ are linearly independent.

Definition 2.1.2. Let $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$ be n elements in a vector space V. The collection of all linear combinations of the form $\sum_{i=1}^{n} \alpha_i \mathbf{u}_i$, where the α_i's are scalars, is called a linear span of $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$ and is denoted by $L(\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n)$. □

It is easy to see from the preceding definition that $L(\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n)$ is a vector subspace of V. This vector subspace is said to be spanned by $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$.

Definition 2.1.3. Let V be a vector space. If there exist linearly independent elements $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$ in V such that $V = L(\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n)$, then $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$ are said to form a basis for V. The number n of elements in this basis is called the dimension of the vector space and is denoted by $\dim V$. □

Note that a basis for a vector space is not unique. However, its dimension is unique. For example, the three vectors $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$ form a basis for R^3. Another basis for R^3 consists of $(1, 1, 0)$, $(1, 0, 1)$, and $(0, 1, 1)$.

If $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$ form a basis for V and if $\mathbf{u}$ is a given element in V, then there exists a unique set of scalars, $\alpha_1, \alpha_2, \ldots, \alpha_n$, such that $\mathbf{u} = \sum_{i=1}^{n} \alpha_i \mathbf{u}_i$. To show this, suppose that there exists another set of scalars, $\beta_1, \beta_2, \ldots, \beta_n$, such that $\mathbf{u} = \sum_{i=1}^{n} \beta_i \mathbf{u}$. Then $\sum_{i=1}^{n} (\alpha_i - \beta_i)\mathbf{u}_i = \mathbf{0}$, which implies that $\alpha_i = \beta_i$ for all i, since the $\mathbf{u}_i$'s are linearly independent.

Let us now check the dimensions of the vector spaces for some of the examples described earlier. For Example 2.1.1, $\dim V = n$. In Example 2.1.2, $\{1, x, x^2, \ldots, x^k\}$ is a basis for V; hence $\dim V = k + 1$. As for Example 2.1.3, $\dim V$ is infinite, since there is no finite set of functions that can span V.

Definition 2.1.4. Let $\mathbf{u}$ and $\mathbf{v}$ be two vectors in R^n. The dot product (also called scalar product or inner product) of $\mathbf{u}$ and $\mathbf{v}$ is a scalar denoted by $\mathbf{u} \cdot \mathbf{v}$ and is given by

$$\mathbf{u} \cdot \mathbf{v} = \sum_{i=1}^{n} u_i v_i,$$

where u_i and v_i are the ith components of $\mathbf{u}$ and $\mathbf{v}$, respectively ($i = 1, 2, \ldots, n$). In particular, if $\mathbf{u} = \mathbf{v}$, then $(\mathbf{u}\cdot\mathbf{u})^{1/2} = (\sum_{i=1}^{n} u_i^2)^{1/2}$ is called the Euclidean norm (or length) of $\mathbf{u}$ and is denoted by $\|\mathbf{u}\|_2$. The dot product of $\mathbf{u}$ and $\mathbf{v}$ is also equal to $\|\mathbf{u}\|_2\|\mathbf{v}\|_2 \cos\theta$, where θ is the angle between $\mathbf{u}$ and $\mathbf{v}$. □

Definition 2.1.5. Two vectors $\mathbf{u}$ and $\mathbf{v}$ in R^n are said to be orthogonal if their dot product is zero. □

Definition 2.1.6. Let U be a vector subspace of R^n. The vectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_m$ form an orthonormal basis for U if they satisfy the following properties:

1. $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_m$ form a basis for U.
2. $\mathbf{e}_i \cdot \mathbf{e}_j = 0$ for all $i \neq j$ $(i, j = 1, 2, \ldots, m)$.
3. $\|\mathbf{e}_i\|_2 = 1$ for $i = 1, 2, \ldots, m$.

Any collection of vectors satisfying just properties 2 and 3 are said to be orthonormal. □

Theorem 2.1.1. Let $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m$ be a basis for a vector subspace U of R^n. Then there exists an orthonormal basis, $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_m$, for U, given by

$$\mathbf{e}_1 = \frac{\mathbf{v}_1}{\|\mathbf{v}_1\|_2}, \qquad \text{where } \mathbf{v}_1 = \mathbf{u}_1,$$

$$\mathbf{e}_2 = \frac{\mathbf{v}_2}{\|\mathbf{v}_2\|_2}, \qquad \text{where } \mathbf{v}_2 = \mathbf{u}_2 - \frac{\mathbf{v}_1 \cdot \mathbf{u}_2}{\|\mathbf{v}_1\|_2^2}\mathbf{v}_1,$$

$$\vdots$$

$$\mathbf{e}_m = \frac{\mathbf{v}_m}{\|\mathbf{v}_m\|_2}, \qquad \text{where } \mathbf{v}_m = \mathbf{u}_m - \sum_{i=1}^{m-1} \frac{\mathbf{v}_i \cdot \mathbf{u}_m}{\|\mathbf{v}_i\|_2^2}\mathbf{v}_i.$$

Proof. See Graybill (1983, Theorem 2.6.5). □

The procedure of constructing an orthonormal basis from any given basis as described in Theorem 2.1.1 is known as the Gram-Schmidt orthonormalization procedure.

Theorem 2.1.2. Let $\mathbf{u}$ and $\mathbf{v}$ be two vectors in R^n. Then:

1. $|\mathbf{u}\cdot\mathbf{v}| \leq \|\mathbf{u}\|_2\|\mathbf{v}\|_2$.
2. $\|\mathbf{u} + \mathbf{v}\|_2 \leq \|\mathbf{u}\|_2 + \|\mathbf{v}\|_2$.

Proof. See Marcus and Minc (1988, Theorem 3.4). □

The inequality in part 1 of Theorem 2.1.2 is known as the *Cauchy–Schwarz inequality*. The one in part 2 is called the *triangle inequality*.

Definition 2.1.7. Let U be a vector subspace of R^n. The orthogonal complement of U, denoted by $U^{\perp}$, is the vector subspace of R^n which consists of all vectors $\mathbf{v}$ such that $\mathbf{u}\cdot\mathbf{v} = 0$ for all $\mathbf{u}$ in U. □

Definition 2.1.8. Let $U_1, U_2, \ldots, U_n$ be vector subspaces of the vector space U. The direct sum of these vector subspaces, denoted by $\oplus_{i=1}^{n} U_i$, consists of all vectors $\mathbf{u}$ that can be uniquely expressed as $\mathbf{u} = \sum_{i=1}^{n} \mathbf{u}_i$, where $\mathbf{u}_i \in U_i$, $i = 1, 2, \ldots, n$. □

Theorem 2.1.3. Let $U_1, U_2, \ldots, U_n$ be vector subspaces of the vector space U. Then:

1. $\oplus_{i=1}^{n} U_i$ is a vector subspace of U.
2. If $U = \oplus_{i=1}^{n} U_i$, then $\cap_{i=1}^{n} U_i$ consists of just the zero element $\mathbf{0}$ of U.
3. $\dim \oplus_{i=1}^{n} U_i = \sum_{i=1}^{n} \dim U_i$.

Proof. The proof is left as an exercise. □

Theorem 2.1.4. Let U be a vector subspace of R^n. Then $R^n = U \oplus U^{\perp}$.

Proof. See Marcus and Minc (1988, Theorem 3.3). □

From Theorem 2.1.4 we conclude that any $\mathbf{v} \in R^n$ can be uniquely written as $\mathbf{v} = \mathbf{v}_1 + \mathbf{v}_2$, where $\mathbf{v}_1 \in U$ and $\mathbf{v}_2 \in U^{\perp}$. In this case, $\mathbf{v}_1$ and $\mathbf{v}_2$ are called the projections of $\mathbf{v}$ on U and $U^{\perp}$, respectively.

2.2. LINEAR TRANSFORMATIONS

Let U and V be two vector spaces. A function $T\colon U \to V$ is called a linear transformation if $T(\alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2) = \alpha_1 T(\mathbf{u}_1) + \alpha_2 T(\mathbf{u}_2)$ for all $\mathbf{u}_1, \mathbf{u}_2$ in U and any scalars α_1 and α_2. For example, let $T\colon R^3 \to R^3$ be defined as

$$T(x_1, x_2, x_3) = (x_1 - x_2, x_1 + x_3, x_3).$$

Then T is a linear transformation, since

$$\begin{aligned} T[\alpha(x_1, x_2, x_3) + \beta(y_1, y_2, y_3)] &= T(\alpha x_1 + \beta y_1, \alpha x_2 + \beta y_2, \alpha x_3 + \beta y_3) \\ &= (\alpha x_1 + \beta y_1 - \alpha x_2 - \beta y_2, \alpha x_1 + \beta y_1 + \alpha x_3 + \beta y_3, \alpha x_3 + \beta y_3) \\ &= \alpha(x_1 - x_2, x_1 + x_3, x_3) + \beta(y_1 - y_2, y_1 + y_3, y_3) \\ &= \alpha T(x_1, x_2, x_3) + \beta T(y_1, y_2, y_3). \end{aligned}$$

We note that the image of U under T, or the range of T, namely $T(U)$, is a vector subspace of V. This is true because if $\mathbf{v}_1, \mathbf{v}_2$ are in $T(U)$, then there exist $\mathbf{u}_1$ and $\mathbf{u}_2$ in U such that $\mathbf{v}_1 = T(\mathbf{u}_1)$ and $\mathbf{v}_2 = T(\mathbf{u}_2)$. Hence, $\mathbf{v}_1 + \mathbf{v}_2 = T(\mathbf{u}_1) + T(\mathbf{u}_2) = T(\mathbf{u}_1 + \mathbf{u}_2)$, which belongs to $T(U)$. Also, if α is a scalar, then $\alpha T(\mathbf{u}) = T(\alpha \mathbf{u}) \in T(U)$ for any $\mathbf{u} \in U$.

Definition 2.2.1. Let $T: U \to V$ be a linear transformation. The kernel of T, denoted by $\ker T$, is the collection of all vectors $\mathbf{u}$ in U such that $T(\mathbf{u}) = \mathbf{0}$, where $\mathbf{0}$ is the zero vector in V. The kernel of T is also called the null space of T.

As an example of a kernel, let $T: R^3 \to R^3$ be defined as $T(x_1, x_2, x_3) = (x_1 - x_2, x_1 - x_3)$. Then

$$\ker T = \{(x_1, x_2, x_3) | x_1 = x_2, x_1 = x_3\}$$

In this case, $\ker T$ consists of all points (x_1, x_2, x_3) in R^3 that lie on a straight line through the origin given by the equations $x_1 = x_2 = x_3$. □

Theorem 2.2.1. Let $T: U \to V$ be a linear transformation. Then we have the following:

1. $\ker T$ is a vector subspace of U.
2. $\dim U = \dim(\ker T) + \dim[T(U)]$.

Proof. Part 1 is left as an exercise. To prove part 2 we consider the following. Let $\dim U = n$, $\dim(\ker T) = p$, and $\dim[T(U)] = q$. Let $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_p$ be a basis for $\ker T$, and $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_q$ be a basis for $T(U)$. Then, there exist vectors $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_q$ in U such that $T(\mathbf{w}_i) = \mathbf{v}_i$ $(i = 1, 2, \ldots, q)$. We need to show that $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_p; \mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_q$ form a basis for U, that is, they are linearly independent and span U.

Suppose that there exist scalars $\alpha_1, \alpha_2, \ldots, \alpha_p; \beta_1, \beta_2, \ldots, \beta_q$ such that

$$\sum_{i=1}^{p} \alpha_i \mathbf{u}_i + \sum_{i=1}^{q} \beta_i \mathbf{w}_i = \mathbf{0}. \tag{2.1}$$

Then

$$\mathbf{0} = T\left(\sum_{i=1}^{p} \alpha_i \mathbf{u}_i + \sum_{i=1}^{q} \beta_i \mathbf{w}_i \right),$$

where $\mathbf{0}$ represents the zero vector in V

$$= \sum_{i=1}^{p} \alpha_i T(\mathbf{u}_i) + \sum_{i=1}^{q} \beta_i T(\mathbf{w}_i)$$

$$= \sum_{i=1}^{q} \beta_i T(\mathbf{w}_i), \qquad \text{since} \quad \mathbf{u}_i \in \ker T,\ i = 1, 2, \ldots, p$$

$$= \sum_{i=1}^{q} \beta_i \mathbf{v}_i.$$

Since the $\mathbf{v}_i$'s are linearly independent, then $\beta_i = 0$ for $i = 1, 2, \ldots, q$. From (2.1) it follows that $\alpha_i = 0$ for $i = 1, 2, \ldots, p$, since the $\mathbf{u}_i$'s are also linearly independent. Thus the vectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_p; \mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_q$ are linearly independent.

Let us now suppose that $\mathbf{u}$ is any vector in U. To show that it belongs to $L(\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_p; \mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_q)$. Let $\mathbf{v} = T(\mathbf{u})$. Then there exist scalars $a_1, a_2, \ldots, a_q$ such that $\mathbf{v} = \sum_{i=1}^{q} a_i \mathbf{v}_i$. It follows that

$$T(\mathbf{u}) = \sum_{i=1}^{q} a_i T(\mathbf{w}_i)$$

$$= T\left(\sum_{i=1}^{q} a_i \mathbf{w}_i \right).$$

Thus,

$$T\left(\mathbf{u} - \sum_{i=1}^{q} a_i \mathbf{w}_i \right) = \mathbf{0},$$

and $\mathbf{u} - \sum_{i=1}^{q} a_i \mathbf{w}_i$ must then belong to $\ker T$. Hence,

$$\mathbf{u} - \sum_{i=1}^{q} a_i \mathbf{w}_i = \sum_{i=1}^{p} b_i \mathbf{u}_i \tag{2.2}$$

for some scalars, $b_1, b_2, \ldots, b_p$. From (2.2) we then have

$$\mathbf{u} = \sum_{i=1}^{p} b_i \mathbf{u}_i + \sum_{i=1}^{q} a_i \mathbf{w}_i,$$

which shows that $\mathbf{u}$ belongs to the linear span of $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_p; \mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_q$. We conclude that these vectors form a basis for U. Hence, $n = p + q$. □

Corollary 2.2.1. $T: U \to V$ is a one-to-one linear transformation if and only if $\dim(\ker T) = 0$.

Proof. If T is a one-to-one linear transformation, then $\ker T$ consists of just one vector, namely, the zero vector. Hence, $\dim(\ker T) = 0$. Vice versa, if $\dim(\ker T) = 0$, or equivalently, if $\ker T$ consists of just the zero vector, then T must be a one-to-one transformation. This is true because if $\mathbf{u}_1$ and $\mathbf{u}_2$ are in U and such that $T(\mathbf{u}_1) = T(\mathbf{u}_2)$, then $T(\mathbf{u}_1 - \mathbf{u}_2) = \mathbf{0}$, which implies that $\mathbf{u}_1 - \mathbf{u}_2 \in \ker T$ and thus $\mathbf{u}_1 - \mathbf{u}_2 = \mathbf{0}$. □

2.3. MATRICES AND DETERMINANTS

Matrix algebra was devised by the English mathematician Arthur Cayley (1821–1895). The use of matrices originated with Cayley in connection with

linear transformations of the form

$$ax_1 + bx_2 = y_1,$$
$$cx_1 + dx_2 = y_2,$$

where a, b, c, and d are scalars. This transformation is completely determined by the square array

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix},$$

which is called a matrix of order 2×2. In general, let $T: U \to V$ be a linear transformation, where U and V are vector spaces of dimensions m and n, respectively. Let $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m$ be a basis for U and $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ be a basis for V. For $i = 1, 2, \ldots, m$, consider $T(\mathbf{u}_i)$, which can be uniquely represented as

$$T(\mathbf{u}_i) = \sum_{j=1}^{n} a_{ij}\mathbf{v}_j, \qquad i = 1, 2, \ldots, m,$$

where the a_{ij}'s are scalars. These scalars completely determine all possible values of T: If $\mathbf{u} \in U$, then $\mathbf{u} = \sum_{i=1}^{m} c_i \mathbf{u}_i$ for some scalars $c_1, c_2, \ldots, c_m$. Then $T(\mathbf{u}) = \sum_{i=1}^{m} c_i T(\mathbf{u}_i) = \sum_{i=1}^{m} c_i (\sum_{j=1}^{n} a_{ij}\mathbf{v}_j)$. By definition, the rectangular array

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

is called a matrix of order $m \times n$, which indicates that $\mathbf{A}$ has m rows and n columns. The a_{ij}'s are called the elements of $\mathbf{A}$. In some cases it is more convenient to represent $\mathbf{A}$ using the notation $\mathbf{A} = (a_{ij})$. In particular, if $m = n$, then $\mathbf{A}$ is called a square matrix. Furthermore, if the off-diagonal elements of a square matrix $\mathbf{A}$ are zero, then $\mathbf{A}$ is called a diagonal matrix and is written as $\mathbf{A} = \text{Diag}(a_{11}, a_{22}, \ldots, a_{nn})$. In this special case, if the diagonal elements are equal to 1, then $\mathbf{A}$ is called the identity matrix and is denoted by $\mathbf{I}_n$ to indicate that it is of order $n \times n$. A matrix of order $m \times 1$ is called a column vector. Likewise, a matrix of order $1 \times n$ is called a row vector.

2.3.1. Basic Operations on Matrices

1. *Equality of Matrices.* Let $\mathbf{A} = (a_{ij})$ and $\mathbf{B} = (b_{ij})$ be two matrices of the same order. Then $\mathbf{A} = \mathbf{B}$ if and only if $a_{ij} = b_{ij}$ for all $i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, n$.

2. *Addition of Matrices*. Let $\mathbf{A} = (a_{ij})$ and $\mathbf{B} = (b_{ij})$ be two matrices of order $m \times n$. Then $\mathbf{A} + \mathbf{B}$ is a matrix $\mathbf{C} = (c_{ij})$ of order $m \times n$ such that $c_{ij} = a_{ij} + b_{ij}$ $(i = 1, 2, \ldots, m;\ j = 1, 2, \ldots, n)$.
3. *Scalar Multiplication.* Let α be a scalar, and $\mathbf{A} = (a_{ij})$ be a matrix of order $m \times n$. Then $\alpha\mathbf{A} = (\alpha a_{ij})$.
4. *The Transpose of a Matrix*. Let $\mathbf{A} = (a_{ij})$ be a matrix of order $m \times n$. The transpose of $\mathbf{A}$, denoted by $\mathbf{A}'$, is a matrix of order $n \times m$ whose rows are the columns of $\mathbf{A}$. For example,

$$\text{if} \quad \mathbf{A} = \begin{bmatrix} 2 & 3 & 1 \\ -1 & 0 & 7 \end{bmatrix}, \quad \text{then} \quad \mathbf{A}' = \begin{bmatrix} 2 & -1 \\ 3 & 0 \\ 1 & 7 \end{bmatrix}.$$

   A matrix $\mathbf{A}$ is symmetric if $\mathbf{A} = \mathbf{A}'$. It is skew-symmetric if $\mathbf{A}' = -\mathbf{A}$. A skew-symmetric matrix must necessarily have zero elements along its diagonal.
5. *Product of Matrices*. Let $\mathbf{A} = (a_{ij})$ and $\mathbf{B} = (b_{ij})$ be matrices of orders $m \times n$ and $n \times p$, respectively. The product $\mathbf{AB}$ is a matrix $\mathbf{C} = (c_{ij})$ of order $m \times p$ such that $c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$ $(i = 1, 2, \ldots, m;\ j = 1, 2, \ldots, p)$. It is to be noted that this product is defined only when the number of columns of $\mathbf{A}$ is equal to the number of rows of $\mathbf{B}$.

   In particular, if $\mathbf{a}$ and $\mathbf{b}$ are column vectors of order $n \times 1$, then their dot product $\mathbf{a} \cdot \mathbf{b}$ can be expressed as a matrix product of the form $\mathbf{a}'\mathbf{b}$ or $\mathbf{b}'\mathbf{a}$.
6. *The Trace of a Matrix*. Let $\mathbf{A} = (a_{ij})$ be a square matrix of order $n \times n$. The trace of $\mathbf{A}$, denoted by $\text{tr}(\mathbf{A})$, is the sum of its diagonal elements, that is,

$$\text{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}.$$

   On the basis of this definition, it is easy to show that if $\mathbf{A}$ and $\mathbf{B}$ are matrices of order $n \times n$, then the following hold: (i) $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$; (ii) $\text{tr}(\mathbf{A} + \mathbf{B}) = \text{tr}(\mathbf{A}) + \text{tr}(\mathbf{B})$.

Definition 2.3.1. Let $\mathbf{A} = (a_{ij})$ be an $m \times n$ matrix. A submatrix $\mathbf{B}$ of $\mathbf{A}$ is a matrix which can be obtained from $\mathbf{A}$ by deleting a certain number of rows and columns.

In particular, if the ith row and jth column of $\mathbf{A}$ that contain the element a_{ij} are deleted, then the resulting matrix is denoted by $\mathbf{M}_{ij}$ $(i = 1, 2, \ldots, m;\ j = 1, 2, \ldots, n)$.

Let us now suppose that $\mathbf{A}$ is a square matrix of order $n \times n$. If rows $i_1, i_2, \ldots, i_p$ and columns $i_1, i_2, \ldots, i_p$ are deleted from $\mathbf{A}$, where $p < n$, then the resulting submatrix is called a principal submatrix of $\mathbf{A}$. In particular, if the deleted rows and columns are the last p rows and the last p columns, respectively, then such a submatrix is called a leading principal submatrix.

Definition 2.3.2. A partitioned matrix is a matrix that consists of several submatrices obtained by drawing horizontal and vertical lines that separate it into groups of rows and columns.

For example, the matrix

$$\mathbf{A} = \left[\begin{array}{c:cc:c} 1 & 0 & 3 & 4 & -5 \\ 6 & 2 & 10 & 5 & 0 \\ \hdashline 3 & 2 & 1 & 0 & 2 \end{array}\right]$$

is partitioned into six submatrices by drawing one horizontal line and two vertical lines as shown above.

Definition 2.3.3. Let $\mathbf{A} = (a_{ij})$ be an $m_1 \times n_1$ matrix and $\mathbf{B}$ be an $m_2 \times n_2$ matrix. The direct (or Kronecker) product of $\mathbf{A}$ and $\mathbf{B}$, denoted by $\mathbf{A} \otimes \mathbf{B}$, is a matrix of order $m_1 m_2 \times n_1 n_2$ defined as a partitioned matrix of the form

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1n_1}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \cdots & a_{2n_1}\mathbf{B} \\ \vdots & \vdots & & \vdots \\ a_{m_1 1}\mathbf{B} & a_{m_2 2}\mathbf{B} & \cdots & a_{m_1 n_1}\mathbf{B} \end{bmatrix}.$$

This matrix can be simplified by writing $\mathbf{A} \otimes \mathbf{B} = [a_{ij}\mathbf{B}]$. □

Properties of the direct product can be found in several matrix algebra books and papers. See, for example, Graybill (1983, Section 8.8), Henderson and Searle (1981), Magnus and Neudecker (1988, Chapter 2), and Searle (1982, Section 10.7). Some of these properties are listed below:

1. $(\mathbf{A} \otimes \mathbf{B})' = \mathbf{A}' \otimes \mathbf{B}'$.
2. $\mathbf{A} \otimes (\mathbf{B} \otimes \mathbf{C}) = (\mathbf{A} \otimes \mathbf{B}) \otimes \mathbf{C}$.
3. $(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = \mathbf{AC} \otimes \mathbf{BD}$, if $\mathbf{AC}$ and $\mathbf{BD}$ are defined.
4. $\operatorname{tr}(\mathbf{A} \otimes \mathbf{B}) = \operatorname{tr}(\mathbf{A})\operatorname{tr}(\mathbf{B})$, if $\mathbf{A}$ and $\mathbf{B}$ are square matrices.

The paper by Henderson, Pukelsheim, and Searle (1983) gives a detailed account of the history associated with direct products.

Definition 2.3.4. Let $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_k$ be matrices of orders $m_i \times n_i$ $(i = 1, 2, \ldots, k)$. The direct sum of these matrices, denoted by $\oplus_{i=1}^{k} \mathbf{A}_i$, is a partitioned matrix of order $(\sum_{i=1}^{k} m_i) \times (\sum_{i=1}^{k} n_i)$ that has the block-diagonal form

$$\bigoplus_{i=1}^{k} \mathbf{A}_i = \operatorname{Diag}(\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_k).$$

The following properties can be easily shown on the basis of the preceding definition:

1. $\oplus_{i=1}^{k}\mathbf{A}_i + \oplus_{i=1}^{k}\mathbf{B}_i = \oplus_{i=1}^{k}(\mathbf{A}_i + \mathbf{B}_i)$, if $\mathbf{A}_i$ and $\mathbf{B}_i$ are of the same order for $i = 1, 2, \ldots, k$.
2. $[\oplus_{i=1}^{k}\mathbf{A}_i][\oplus_{i=1}^{k}\mathbf{B}_i] = \oplus_{i=1}^{k}\mathbf{A}_i\mathbf{B}_i$, if $\mathbf{A}_i\mathbf{B}_i$ is defined for $i = 1, 2, \ldots, k$.
3. $[\oplus_{i=1}^{k}\mathbf{A}_i]' = \oplus_{i=1}^{k}\mathbf{A}_i'$.
4. $\mathrm{tr}(\oplus_{i=1}^{k}\mathbf{A}_i) = \sum_{i=1}^{k}\mathrm{tr}(\mathbf{A}_i)$. □

Definition 2.3.5. Let $\mathbf{A} = (a_{ij})$ be a square matrix of order $n \times n$. The determinant of $\mathbf{A}$, denoted by $\det(\mathbf{A})$, is a scalar quantity that can be computed iteratively as

$$\det(\mathbf{A}) = \sum_{j=1}^{n} (-1)^{j+1} a_{1j} \det(\mathbf{M}_{1j}), \tag{2.3}$$

where $\mathbf{M}_{1j}$ is a submatrix of $\mathbf{A}$ obtained by deleting row 1 and column j ($j = 1, 2, \ldots, n$). For each j, the determinant of $\mathbf{M}_{1j}$ is obtained in terms of determinants of matrices of order $(n-2) \times (n-2)$ using a formula similar to (2.3). This process is repeated several times until the matrices on the right-hand side of (2.3) become of order 2×2. The determinant of a 2×2 matrix such as $\mathbf{b} = (b_{ij})$ is given by $\det(\mathbf{B}) = b_{11}b_{22} - b_{12}b_{21}$. Thus by an iterative application of formula (2.3), the value of $\det(\mathbf{A})$ can be fully determined. For example, let $\mathbf{A}$ be the matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & -1 \\ 5 & 0 & 3 \\ 1 & 2 & 1 \end{bmatrix}.$$

Then $\det(\mathbf{A}) = \det(\mathbf{A}_1) - 2\det(\mathbf{A}_2) - \det(\mathbf{A}_3)$, where $\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3$ are 2×2 submatrices, namely

$$\mathbf{A}_1 = \begin{bmatrix} 0 & 3 \\ 2 & 1 \end{bmatrix}, \qquad \mathbf{A}_2 = \begin{bmatrix} 5 & 3 \\ 1 & 1 \end{bmatrix}, \qquad \mathbf{A}_3 = \begin{bmatrix} 5 & 0 \\ 1 & 2 \end{bmatrix}.$$

It follows that $\det(\mathbf{A}) = -6 - 2(2) - 10 = -20$. □

Definition 2.3.6. Let $\mathbf{A} = (a_{ij})$ be a square matrix order of $n \times n$. The determinant of $\mathbf{M}_{ij}$, the submatrix obtained by deleting row i and column j, is called a minor of $\mathbf{A}$ of order $n - 1$. The quantity $(-1)^{i+j}\det(\mathbf{M}_{ij})$ is called a cofactor of the corresponding (i, j)th element of $\mathbf{A}$. More generally, if $\mathbf{A}$ is an $m \times n$ matrix and if we strike out all but p rows and the same number of columns from $\mathbf{A}$, where $p \leq \min(m, n)$, then the determinant of the resulting submatrix is called a minor of $\mathbf{A}$ of order p.

The determinant of a principal submatrix of a square matrix **A** is called a principal minor. If, however, we have a leading principal submatrix, then its determinant is called a leading principal minor. □

NOTE 2.3.1. The determinant of a matrix **A** is defined only when **A** is a square matrix.

NOTE 2.3.2. The expansion of $\det(\mathbf{A})$ in (2.3) was carried out by multiplying the elements of the first row of **A** by their corresponding cofactors and then summing over $j\ (=1,2,\ldots,n)$. The same value of $\det(\mathbf{A})$ could have also been obtained by similar expansions according to the elements of any row of **A** (instead of the first row), or any column of **A**. Thus if $\mathbf{M}_{ij}$ is a submatrix of **A** obtained by deleting row i and column j, then $\det(\mathbf{A})$ can be obtained by using any of the following expansions:

$$\text{By row } i: \qquad \det(\mathbf{A}) = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det(\mathbf{M}_{ij}), \qquad i = 1,2,\ldots,n.$$

$$\text{By column } j: \qquad \det(\mathbf{A}) = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} \det(\mathbf{M}_{ij}), \qquad j = 1,2,\ldots,n.$$

NOTE 2.3.3. Some of the properties of determinants are the following:

- **i.** $\det(\mathbf{AB}) = \det(\mathbf{A})\det(\mathbf{B})$, if **A** and **B** are $n \times n$ matrices.
- **ii.** If $\mathbf{A}'$ is the transpose of **A**, then $\det(\mathbf{A}') = \det(\mathbf{A})$.
- **iii.** If **A** is an $n \times n$ matrix and α is a scalar, then $\det(\alpha\mathbf{A}) = \alpha^n \det(\mathbf{A})$.
- **iv.** If any two rows (or columns) of **A** are identical, then $\det(\mathbf{A}) = 0$.
- **v.** If any two rows (or columns) of **A** are interchanged, then $\det(\mathbf{A})$ is multiplied by -1.
- **vi.** If $\det(\mathbf{A}) = 0$, then **A** is called a singular matrix. Otherwise, **A** is a nonsingular matrix.
- **vii.** If **A** and **B** are matrices of orders $m \times m$ and $n \times n$, respectively, then the following hold: (a) $\det(\mathbf{A} \otimes \mathbf{B}) = [\det(\mathbf{A})]^n[\det(\mathbf{B})]^m$; (b) $\det(\mathbf{A} \oplus \mathbf{B}) = [\det(\mathbf{A})][\det(\mathbf{B})]$.

NOTE 2.3.4. The history of determinants dates back to the fourteenth century. According to Smith (1958, page 273), the Chinese had some knowledge of determinants as early as about 1300 A.D. Smith (1958, page 440) also reported that the Japanese mathematician Seki Kõwa (1642–1708) had discovered the expansion of a determinant in solving simultaneous equations. In the West, the theory of determinants is believed to have originated with the German mathematician Gottfried Leibniz (1646–1716) in 1693, ten years

after the work of Seki Kõwa. However, the actual development of the theory of determinants did not begin until the publication of a book by Gabriel Cramer (1704–1752) (see Price, 1947, page 85) in 1750. Other mathematicians who contributed to this theory include Alexandre Vandermonde (1735–1796), Pierre-Simon Laplace (1749–1827), Carl Gauss (1777–1855), and Augustin-Louis Cauchy (1789–1857). Arthur Cayley (1821–1895) is credited with having been the first to introduce the common present-day notation of vertical bars enclosing a square matrix. For more interesting facts about the history of determinants, the reader is advised to read the article by Price (1947).

2.3.2. The Rank of a Matrix

Let $\mathbf{A} = (a_{ij})$ be a matrix of order $m \times n$. Let $\mathbf{u}'_1, \mathbf{u}'_2, \ldots, \mathbf{u}'_m$ denote the row vectors of $\mathbf{A}$, and let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ denote its column vectors. Consider the linear spans of the row and column vectors, namely, $V_1 = L(\mathbf{u}'_1, \mathbf{u}'_2, \ldots, \mathbf{u}'_m)$, $V_2 = L(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$, respectively.

Theorem 2.3.1. The vector spaces V_1 and V_2 have the same dimension.

Proof. See Lancaster (1969, Theorem 1.15.1), or Searle (1982, Section 6.6). □

Thus, for any matrix $\mathbf{A}$, the number of linearly independent rows is the same as the number of linearly independent columns.

Definition 2.3.7. The rank of a matrix $\mathbf{A}$ is the number of its linearly independent rows (or columns). The rank of $\mathbf{A}$ is denoted by $r(\mathbf{A})$. □

Theorem 2.3.2. If a matrix $\mathbf{A}$ has a nonzero minor of order r, and if all minors of order $r+1$ and higher (if they exist) are zero, then $\mathbf{A}$ has rank r.

Proof. See Lancaster (1969, Lemma 1, Section 1.15). □

For example, if $\mathbf{A}$ is the matrix

$$\mathbf{A} = \begin{bmatrix} 2 & 3 & -1 \\ 0 & 1 & 2 \\ 2 & 4 & 1 \end{bmatrix},$$

then $r(\mathbf{A}) = 2$. This is because $\det(\mathbf{A}) = 0$ and at least one minor of order 2 is different from zero.

There are several properties associated with the rank of a matrix. Some of these properties are the following:

1. $r(\mathbf{A}) = r(\mathbf{A}')$.
2. The rank of $\mathbf{A}$ is unchanged if $\mathbf{A}$ is multiplied by a nonsingular matrix. Thus if $\mathbf{A}$ is an $m \times n$ matrix and $\mathbf{P}$ is an $n \times n$ nonsingular matrix, then $r(\mathbf{A}) = r(\mathbf{AP})$.
3. $r(\mathbf{A}) = r(\mathbf{AA}') = r(\mathbf{A}'\mathbf{A})$.
4. If the matrix $\mathbf{A}$ is partitioned as $\mathbf{A} = [\mathbf{A}_1 : \mathbf{A}_2]$, where $\mathbf{A}_1$ and $\mathbf{A}_2$ are submatrices of the same order, then $r(\mathbf{A}_1 + \mathbf{A}_2) \le r(\mathbf{A}) \le r(\mathbf{A}_1) + r(\mathbf{A}_2)$. More generally, if the matrices $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_k$ are of the same order and if $\mathbf{A}$ is partitioned as $\mathbf{A} = [\mathbf{A}_1 : \mathbf{A}_2 : \cdots : \mathbf{A}_k]$, then

$$r\left(\sum_{i=1}^{k} \mathbf{A}_i\right) \le r(\mathbf{A}) \le \sum_{i=1}^{k} r(\mathbf{A}_i).$$

5. If the product $\mathbf{AB}$ is defined, then $r(\mathbf{A}) + r(\mathbf{B}) - n \le r(\mathbf{AB}) \le \min\{r(\mathbf{A}), r(\mathbf{B})\}$, where n is the number of columns of $\mathbf{A}$ (or the number of rows of $\mathbf{B}$).
6. $r(\mathbf{A} \otimes \mathbf{B}) = r(\mathbf{A})r(\mathbf{B})$.
7. $r(\mathbf{A} \oplus \mathbf{B}) = r(\mathbf{A}) + r(\mathbf{B})$.

Definition 2.3.8. Let $\mathbf{A}$ be a matrix of order $m \times n$ and rank r. Then we have the following:

1. $\mathbf{A}$ is said to have a full row rank if $r = m < n$.
2. $\mathbf{A}$ is said to have a full column rank if $r = n < m$.
3. $\mathbf{A}$ is of full rank if $r = m = n$. In this case, $\det(\mathbf{A}) \neq 0$, that is, $\mathbf{A}$ is a nonsingular matrix. □

2.3.3. The Inverse of a Matrix

Let $\mathbf{A} = (a_{ij})$ be a nonsingular matrix of order $n \times n$. The inverse of $\mathbf{A}$, denoted by $\mathbf{A}^{-1}$, is an $n \times n$ matrix that satisfies the condition $\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_n$.

The inverse of $\mathbf{A}$ can be computed as follows: Let c_{ij} be the cofactor of a_{ij} (see Definition 2.3.6). Define the matrix $\mathbf{C}$ as $\mathbf{C} = (c_{ij})$. The transpose of $\mathbf{C}$ is called the adjugate or adjoint of $\mathbf{A}$ and is denoted by $\operatorname{adj}\mathbf{A}$. The inverse of $\mathbf{A}$ is then given by

$$\mathbf{A}^{-1} = \frac{\operatorname{adj}\mathbf{A}}{\det(\mathbf{A})}.$$

It can be verified that

$$\mathbf{A}\left[\frac{\operatorname{adj}\mathbf{A}}{\det(\mathbf{A})}\right] = \left[\frac{\operatorname{adj}\mathbf{A}}{\det(\mathbf{A})}\right]\mathbf{A} = \mathbf{I}_n.$$

For example, if $\mathbf{A}$ is the matrix

$$\mathbf{A} = \begin{bmatrix} 2 & 0 & 1 \\ -3 & 2 & 0 \\ 2 & 1 & 1 \end{bmatrix},$$

then $\det(\mathbf{A}) = -3$, and

$$\operatorname{adj}\mathbf{A} = \begin{bmatrix} 2 & 1 & -2 \\ 3 & 0 & -3 \\ -7 & -2 & 4 \end{bmatrix}.$$

Hence,

$$\mathbf{A}^{-1} = \begin{bmatrix} -\frac{2}{3} & -\frac{1}{3} & \frac{2}{3} \\ -1 & 0 & 1 \\ \frac{7}{3} & \frac{2}{3} & -\frac{4}{3} \end{bmatrix}.$$

Some properties of the inverse operation are given below:

1. $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.
2. $(\mathbf{A}')^{-1} = (\mathbf{A}^{-1})'$.
3. $\det(\mathbf{A}^{-1}) = 1/\det(\mathbf{A})$.
4. $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$.
5. $(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}$.
6. $(\mathbf{A} \oplus \mathbf{B})^{-1} = \mathbf{A}^{-1} \oplus \mathbf{B}^{-1}$.
7. If $\mathbf{A}$ is partitioned as

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix},$$

where $\mathbf{A}_{ij}$ is of order $n_i \times n_j$ $(i, j = 1, 2)$, then

$$\det(\mathbf{A}) = \begin{cases} \det(\mathbf{A}_{11}) \cdot \det\left(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\right) & \text{if } \mathbf{A}_{11} \text{ is nonsingular,} \\ \det(\mathbf{A}_{22}) \cdot \det\left(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\right) & \text{if } \mathbf{A}_{22} \text{ is nonsingular.} \end{cases}$$

The inverse of $\mathbf{A}$ is partitioned as

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{bmatrix},$$

where

$$\mathbf{B}_{11} = \left(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\right)^{-1},$$

$$\mathbf{B}_{12} = -\mathbf{B}_{11}\mathbf{A}_{12}\mathbf{A}_{22}^{-1},$$

$$\mathbf{B}_{21} = -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{B}_{11},$$

$$\mathbf{B}_{22} = \mathbf{A}_{22}^{-1} + \mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{B}_{11}\mathbf{A}_{12}\mathbf{A}_{22}^{-1}.$$

2.3.4. Generalized Inverse of a Matrix

This inverse represents a more general concept than the one discussed in the previous section. Let $\mathbf{A}$ be a matrix of order $m \times n$. Then, a generalized inverse of $\mathbf{A}$, denoted by $\mathbf{A}^-$, is a matrix of order $n \times m$ that satisfies the condition

$$\mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A}. \tag{2.4}$$

Note that $\mathbf{A}^-$ is defined even if $\mathbf{A}$ is not a square matrix. If $\mathbf{A}$ is a square matrix, it does not have to be nonsingular. Furthermore, condition (2.4) can be satisfied by infinitely many matrices (see, for example, Searle, 1982, Chapter 8). If $\mathbf{A}$ is nonsingular, then (2.4) is satisfied by only $\mathbf{A}^{-1}$. Thus $\mathbf{A}^{-1}$ is a special case of $\mathbf{A}^-$.

Theorem 2.3.3.

1. If $\mathbf{A}$ is a symmetric matrix, then $\mathbf{A}^-$ can be chosen to be symmetric.
2. $\mathbf{A}(\mathbf{A}'\mathbf{A})^-\mathbf{A}'\mathbf{A} = \mathbf{A}$ for any matrix $\mathbf{A}$.
3. $\mathbf{A}(\mathbf{A}'\mathbf{A})^-\mathbf{A}'$ is invariant to the choice of a generalized inverse of $\mathbf{A}'\mathbf{A}$.

Proof. See Searle (1982, pages 221–222). □

2.3.5. Eigenvalues and Eigenvectors of a Matrix

Let $\mathbf{A}$ be a square matrix of order $n \times n$. By definition, a scalar λ is said to be an eigenvalue (or characteristic root) of $\mathbf{A}$ if $\mathbf{A} - \lambda\mathbf{I}_n$ is a singular matrix, that is,

$$\det(\mathbf{A} - \lambda\mathbf{I}_n) = 0. \tag{2.5}$$

Thus an eigenvalue of $\mathbf{A}$ satisfies a polynomial equation of degree n called the characteristic equation of $\mathbf{A}$. If λ is a multiple solution (or root) of equation (2.5), that is, (2.5) has several roots, say m, that are equal to λ, then λ is said to be an eigenvalue of multiplicity m.

Since $r(\mathbf{A} - \lambda\mathbf{I}_n) < n$ by the fact that $\mathbf{A} - \lambda\mathbf{I}_n$ is singular, the columns of $\mathbf{A} - \lambda\mathbf{I}_n$ must be linearly related. Hence, there exists a nonzero vector $\mathbf{v}$ such that

$$(\mathbf{A} - \lambda\mathbf{I}_n)\mathbf{v} = \mathbf{0}, \tag{2.6}$$

or equivalently,

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v}. \tag{2.7}$$

A vector satisfying (2.7) is called an eigenvector (or a characteristic vector) corresponding to the eigenvalue λ. From (2.7) we note that the linear transformation of $\mathbf{v}$ by the matrix $\mathbf{A}$ is a scalar multiple of $\mathbf{v}$.

The following theorems describe certain properties associated with eigenvalues and eigenvectors. The proofs of these theorems can be found in standard matrix algebra books (see the annotated bibliography).

Theorem 2.3.4. A square matrix $\mathbf{A}$ is singular if and only if at least one of its eigenvalues is equal to zero. In particular, if $\mathbf{A}$ is symmetric, then its rank is equal to the number of its nonzero eigenvalues.

Theorem 2.3.5. The eigenvalues of a symmetric matrix are real.

Theorem 2.3.6. Let $\mathbf{A}$ be a square matrix, and let $\lambda_1, \lambda_2, \ldots, \lambda_k$ denote its distinct eigenvalues. If $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$ are eigenvectors of $\mathbf{A}$ corresponding to $\lambda_1, \lambda_2, \ldots, \lambda_k$, respectively, then $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$ are linearly independent. In particular, if $\mathbf{A}$ is symmetric, then $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$ are orthogonal to one another, that is, $\mathbf{v}_i'\mathbf{v}_j = 0$ for $i \neq j$ $(i, j = 1, 2, \ldots, k)$.

Theorem 2.3.7. Let $\mathbf{A}$ and $\mathbf{B}$ be two matrices of orders $m \times m$ and $n \times n$, respectively. Let $\lambda_1, \lambda_2, \ldots, \lambda_m$ be the eigenvalues of $\mathbf{A}$, and $\upsilon_1, \upsilon_2, \ldots, \upsilon_n$ be the eigenvalues of $\mathbf{B}$. Then we have the following:

1. The eigenvalues of $\mathbf{A} \otimes \mathbf{B}$ are of the form $\lambda_i\upsilon_j$ $(i = 1, 2, \ldots, m;\ j = 1, 2, \ldots, n)$.
2. The eigenvalues of $\mathbf{A} \oplus \mathbf{B}$ are $\lambda_1, \lambda_2, \ldots, \lambda_m;\ \upsilon_1, \upsilon_2, \ldots, \upsilon_n$.

Theorem 2.3.8. Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the eigenvalues of a matrix $\mathbf{A}$ of order $n \times n$. Then the following hold:

1. $\operatorname{tr}(\mathbf{A}) = \sum_{i=1}^{n} \lambda_i$.
2. $\det(\mathbf{A}) = \prod_{i=1}^{n} \lambda_i$.

Theorem 2.3.9. Let $\mathbf{A}$ and $\mathbf{B}$ be two matrices of orders $m \times n$ and $n \times m$ $(n \geq m)$, respectively. The nonzero eigenvalues of $\mathbf{BA}$ are the same as those of $\mathbf{AB}$.

2.3.6. Some Special Matrices

1. The vector $\mathbf{1}_n$ is a column vector of ones of order $n \times 1$.
2. The matrix $\mathbf{J}_n$ is a matrix of ones of order $n \times n$.
3. *Idempotent Matrix*. A square matrix $\mathbf{A}$ for which $\mathbf{A}^2 = \mathbf{A}$ is called an idempotent matrix. For example, the matrix $\mathbf{A} = \mathbf{I}_n - (1/n)\mathbf{J}_n$ is idempotent of order $n \times n$. The eigenvalues of an idempotent matrix are equal to zeros and ones. It follows from Theorem 2.3.8 that the rank of an idempotent matrix, which is the same as the number of eigenvalues that are equal to 1, is also equal to its trace. Idempotent matrices are used in many applications in statistics (see Section 2.4).
4. *Orthogonal Matrix*. A square matrix $\mathbf{A}$ is orthogonal if $\mathbf{A}'\mathbf{A} = \mathbf{I}$. From this definition it follows that (i) $\mathbf{A}$ is orthogonal if and only if $\mathbf{A}' = \mathbf{A}^{-1}$; (ii) $|\det(\mathbf{A})| = 1$. A special orthogonal matrix is the Householder matrix, which is a symmetric matrix of the form

$$\mathbf{H} = \mathbf{I} - 2\mathbf{u}\mathbf{u}'/\mathbf{u}'\mathbf{u},$$

where $\mathbf{u}$ is a nonzero vector. Orthogonal matrices occur in many applications of matrix algebra and play an important role in statistics, as will be seen in Section 2.4.

2.3.7. The Diagonalization of a Matrix

Theorem 2.3.10 (The Spectral Decomposition Theorem). Let $\mathbf{A}$ be a symmetric matrix of order $n \times n$. There exists an orthogonal matrix $\mathbf{P}$ such that $\mathbf{A} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}'$, where $\mathbf{\Lambda} = \mathrm{Diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ is a diagonal matrix whose diagonal elements are the eigenvalues of $\mathbf{A}$. The columns of $\mathbf{P}$ are the corresponding orthonormal eigenvectors of $\mathbf{A}$.

Proof. See Basilevsky (1983, Theorem 5.8, page 200). □

If $\mathbf{P}$ is partitioned as $\mathbf{P} = [\mathbf{p}_1 : \mathbf{p}_2 : \cdots : \mathbf{p}_n]$, where $\mathbf{p}_i$ is an eigenvector of $\mathbf{A}$ with eigenvalue λ_i $(i = 1, 2, \ldots, n)$, then $\mathbf{A}$ can be written as

$$\mathbf{A} = \sum_{i=1}^{n} \lambda_i \mathbf{p}_i \mathbf{p}_i'.$$

For example, if

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & -2 \\ 0 & 0 & 0 \\ -2 & 0 & 4 \end{bmatrix},$$

then $\mathbf{A}$ has two distinct eigenvalues, $\lambda_1 = 0$ of multiplicity 2 and $\lambda_2 = 5$. For $\lambda_1 = 0$ we have two orthonormal eigenvectors, $\mathbf{p}_1 = (2, 0, 1)'/\sqrt{5}$ and $\mathbf{p}_2 = (0, 1, 0)'$. Note that $\mathbf{p}_1$ and $\mathbf{p}_2$ span the kernel (null space) of the linear transformation represented by $\mathbf{A}$. For $\lambda_2 = 5$ we have the normal eigenvector $\mathbf{p}_3 = (1, 0, -2)'/\sqrt{5}$, which is orthogonal to both $\mathbf{p}_1$ and $\mathbf{p}_2$. Hence, $\mathbf{P}$ and $\mathbf{\Lambda}$ in Theorem 2.3.10 for the matrix $\mathbf{A}$ are

$$\mathbf{P} = \begin{bmatrix} \frac{2}{\sqrt{5}} & 0 & \frac{1}{\sqrt{5}} \\ 0 & 1 & 0 \\ \frac{1}{\sqrt{5}} & 0 & \frac{-2}{\sqrt{5}} \end{bmatrix},$$

$$\mathbf{\Lambda} = \operatorname{Diag}(0, 0, 5).$$

The next theorem gives a more general form of the spectral decomposition theorem.

Theorem 2.3.11 (The Singular-Value Decomposition Theorem). Let $\mathbf{A}$ be a matrix of order $m \times n$ $(m \leq n)$ and rank r. There exist orthogonal matrices $\mathbf{P}$ and $\mathbf{Q}$ such that $\mathbf{A} = \mathbf{P}[\mathbf{D} : \mathbf{0}]\mathbf{Q}'$, where $\mathbf{D} = \operatorname{Diag}(\lambda_1, \lambda_2, \ldots, \lambda_m)$ is a diagonal matrix with nonnegative diagonal elements called the singular values of $\mathbf{A}$, and $\mathbf{0}$ is a zero matrix of order $m \times (n - m)$. The diagonal elements of $\mathbf{D}$ are the square roots of the eigenvalues of $\mathbf{AA}'$.

Proof. See, for example, Searle (1982, pages 316–317). □

2.3.8. Quadratic Forms

Let $\mathbf{A} = (a_{ij})$ be a symmetric matrix of order $n \times n$, and let $\mathbf{x} = (x_1, x_2, \ldots, x_n)'$ be a column vector of order $n \times 1$. The function

$$\begin{aligned} q(\mathbf{x}) &= \mathbf{x}'\mathbf{A}\mathbf{x} \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j \end{aligned}$$

is called a quadratic form in $\mathbf{x}$.

A quadratic form $\mathbf{x}'\mathbf{A}\mathbf{x}$ is said to be the following:

1. Positive definite if $\mathbf{x}'\mathbf{A}\mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$ and is zero only if $\mathbf{x} = \mathbf{0}$.
2. Positive semidefinite if $\mathbf{x}'\mathbf{A}\mathbf{x} \geq 0$ for all $\mathbf{x}$ and $\mathbf{x}'\mathbf{A}\mathbf{x} = 0$ for at least one nonzero value of $\mathbf{x}$.
3. Nonnegative definite if $\mathbf{A}$ is either positive definite or positive semidefinite.

Theorem 2.3.12. Let $\mathbf{A} = (a_{ij})$ be a symmetric matrix of order $n \times n$. Then $\mathbf{A}$ is positive definite if and only if either of the following two conditions is satisfied:

1. The eigenvalues of $\mathbf{A}$ are all positive.
2. The leading principal minors of $\mathbf{A}$ are all positive, that is,

$$a_{11} > 0, \qquad \det\left(\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}\right) > 0, \ldots, \qquad \det(\mathbf{A}) > 0.$$

Proof. The proof of part 1 follows directly from the spectral decomposition theorem. For the proof of part 2, see Lancaster (1969, Theorem 2.14.4). □

Theorem 2.3.13. Let $\mathbf{A} = (a_{ij})$ be a symmetric matrix of order $n \times n$. Then $\mathbf{A}$ is positive semidefinite if and only if its eigenvalues are nonnegative with at least one of them equal to zero.

Proof. See Basilevsky (1983, Theorem 5.10, page 203). □

2.3.9. The Simultaneous Diagonalization of Matrices

By simultaneous diagonalization we mean finding a matrix, say $\mathbf{Q}$, that can reduce several square matrices to a diagonal form. In many situations there may be a need to diagonalize several matrices simultaneously. This occurs frequently in statistics, particularly in analysis of variance.

The proofs of the following theorems can be found in Graybill (1983, Chapter 12).

Theorem 2.3.14. Let $\mathbf{A}$ and $\mathbf{B}$ be symmetric matrices of order $n \times n$.

1. If $\mathbf{A}$ is positive definite, then there exists a nonsingular matrix $\mathbf{Q}$ such that $\mathbf{Q}'\mathbf{A}\mathbf{Q} = \mathbf{I}_n$ and $\mathbf{Q}'\mathbf{B}\mathbf{Q} = \mathbf{D}$, where $\mathbf{D}$ is a diagonal matrix whose diagonal elements are the roots of the polynomial equation $\det(\mathbf{B} - \lambda\mathbf{A}) = 0$.

2. If **A** and **B** are positive semidefinite, then there exists a nonsingular matrix **Q** such that

$$\mathbf{Q}'\mathbf{A}\mathbf{Q} = \mathbf{D}_1,$$
$$\mathbf{Q}'\mathbf{B}\mathbf{Q} = \mathbf{D}_2,$$

where $\mathbf{D}_1$ and $\mathbf{D}_2$ are diagonal matrices (for a detailed proof of this result, see Newcomb, 1960).

Theorem 2.3.15. Let $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_k$ be symmetric matrices of order $n \times n$. Then there exists an orthogonal matrix **P** such that

$$\mathbf{A}_i = \mathbf{P}\mathbf{\Lambda}_i\mathbf{P}', \qquad i = 1, 2, \ldots, k,$$

where $\mathbf{\Lambda}_i$ is a diagonal matrix, if and only if $\mathbf{A}_i\mathbf{A}_j = \mathbf{A}_j\mathbf{A}_i$ for all $i \neq j$ $(i, j = 1, 2, \ldots, k)$.

2.3.10. Bounds on Eigenvalues

Let **A** be a symmetric matrix of order $n \times n$. We denote the ith eigenvalue of **A** by $e_i(\mathbf{A})$, $i = 1, 2, \ldots, n$. The smallest and largest eigenvalues of **A** are denoted by $e_{\min}(\mathbf{A})$ and $e_{\max}(\mathbf{A})$, respectively.

Theorem 2.3.16. $e_{\min}(\mathbf{A}) \leq \mathbf{x}'\mathbf{A}\mathbf{x}/\mathbf{x}'\mathbf{x} \leq e_{\max}(\mathbf{A})$.

Proof. This follows directly from the spectral decomposition theorem. □

The ratio $\mathbf{x}'\mathbf{A}\mathbf{x}/\mathbf{x}'\mathbf{x}$ is called Rayleigh's quotient for **A**. The lower and upper bounds in Theorem 2.3.16 can be achieved by choosing **x** to be an eigenvector associated with $e_{\min}(\mathbf{A})$ and $e_{\max}(\mathbf{A})$, respectively. Thus Theorem 2.3.16 implies that

$$\inf_{\mathbf{x} \neq \mathbf{0}} \left[\frac{\mathbf{x}'\mathbf{A}\mathbf{x}}{\mathbf{x}'\mathbf{x}} \right] = e_{\min}(\mathbf{A}), \tag{2.8}$$

$$\sup_{\mathbf{x} \neq \mathbf{0}} \left[\frac{\mathbf{x}'\mathbf{A}\mathbf{x}}{\mathbf{x}'\mathbf{x}} \right] = e_{\max}(\mathbf{A}). \tag{2.9}$$

Theorem 2.3.17. If **A** is a symmetric matrix and **B** is a positive definite matrix, both of order $n \times n$, then

$$e_{\min}(\mathbf{B}^{-1}\mathbf{A}) \leq \frac{\mathbf{x}'\mathbf{A}\mathbf{x}}{\mathbf{x}'\mathbf{B}\mathbf{x}} \leq e_{\max}(\mathbf{B}^{-1}\mathbf{A})$$

Proof. The proof is left to the reader. □

Note that the above lower and upper bounds are equal to the infimum and supremum, respectively, of the ratio $\mathbf{x}'\mathbf{A}\mathbf{x}/\mathbf{x}'\mathbf{B}\mathbf{x}$ for $\mathbf{x} \neq \mathbf{0}$.

Theorem 2.3.18. If $\mathbf{A}$ is a positive semidefinite matrix and $\mathbf{B}$ is a positive definite matrix, both of order $n \times n$, then for any i $(i = 1, 2, \ldots, n)$,

$$e_i(\mathbf{A})e_{\min}(\mathbf{B}) \leq e_i(\mathbf{A}\mathbf{B}) \leq e_i(\mathbf{A})e_{\max}(\mathbf{B}). \tag{2.10}$$

Furthermore, if $\mathbf{A}$ is positive definite, then for any i $(i = 1, 2, \ldots, n)$,

$$\frac{e_i^2(\mathbf{A}\mathbf{B})}{e_{\max}(\mathbf{A})e_{\max}(\mathbf{B})} \leq e_i(\mathbf{A})e_i(\mathbf{B}) \leq \frac{e_i^2(\mathbf{A}\mathbf{B})}{e_{\min}(\mathbf{A})e_{\min}(\mathbf{B})}$$

Proof. See Anderson and Gupta (1963, Corollary 2.2.1). □

A special case of the double inequality in (2.10) is

$$e_{\min}(\mathbf{A})e_{\min}(\mathbf{B}) \leq e_i(\mathbf{A}\mathbf{B}) \leq e_{\max}(\mathbf{A})e_{\max}(\mathbf{B}),$$

for all i $(i = 1, 2, \ldots, n)$.

Theorem 2.3.19. Let $\mathbf{A}$ and $\mathbf{B}$ be symmetric matrices of order $n \times n$. Then, the following hold:

1. $e_i(\mathbf{A}) \leq e_i(\mathbf{A} + \mathbf{B})$, $i = 1, 2, \ldots, n$, if $\mathbf{B}$ is nonnegative definite.
2. $e_i(\mathbf{A}) < e_i(\mathbf{A} + \mathbf{B})$, $i = 1, 2, \ldots, n$, if $\mathbf{B}$ is positive definite.

Proof. See Bellman (1970, Theorem 3, page 117). □

Theorem 2.3.20 (Schur's Theorem). Let $\mathbf{A} = (a_{ij})$ be a symmetric matrix of order $n \times n$, and let $\|\mathbf{A}\|_2$ denote its Euclidean norm, defined as

$$\|\mathbf{A}\|_2 = \left(\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}^2 \right)^{1/2}.$$

Then

$$\sum_{i=1}^{n} e_i^2(\mathbf{A}) = \|\mathbf{A}\|_2^2.$$

Proof. See Lancaster (1969, Theorem 7.3.1). □

Since $\|\mathbf{A}\|_2 \leq n \max_{i,j} |a_{ij}|$, then from Theorem 2.3.20 we conclude that

$$|e_{\max}(\mathbf{A})| \leq n \max_{i,j} |a_{ij}|.$$

Theorem 2.3.21. Let $\mathbf{A}$ be a symmetric matrix of order $n \times n$, and let m and s be defined as

$$m = \frac{\operatorname{tr}(\mathbf{A})}{n}, \qquad s = \left(\frac{\operatorname{tr}(\mathbf{A}^2)}{n} - m^2 \right)^{1/2}.$$

Then

$$m - s(n-1)^{1/2} \le e_{\min}(\mathbf{A}) \le m - \frac{s}{(n-1)^{1/2}},$$

$$m + \frac{s}{(n-1)^{1/2}} \le e_{\max}(\mathbf{A}) \le m + s(n-1)^{1/2},$$

$$e_{\max}(\mathbf{A}) - e_{\min}(\mathbf{A}) \le s(2n)^{1/2}.$$

Proof. See Wolkowicz and Styan (1980, Theorems 2.1 and 2.5). □

2.4. APPLICATIONS OF MATRICES IN STATISTICS

The use of matrix algebra is quite prevalent in statistics. In fact, in the areas of experimental design, linear models, and multivariate analysis, matrix algebra is considered the most frequently used branch of mathematics. Applications of matrices in these areas are well documented in several books, for example, Basilevsky (1983), Graybill (1983), Magnus and Neudecker (1988), and Searle (1982). We shall therefore not attempt to duplicate the material given in these books.

Let us consider the following applications:

2.4.1. The Analysis of the Balanced Mixed Model

In analysis of variance, a linear model associated with a given experimental situation is said to be balanced if the numbers of observations in the subclasses of the data are the same. For example, the two-way crossed-classification model with interaction,

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}, \tag{2.11}$$

$i = 1, 2, \ldots, a;\ j = 1, 2, \ldots, b;\ k = 1, 2, \ldots, n,$ is balanced, since there are n observations for each combination of i and j. Here, α_i and β_j represent the main effects of the factors under consideration, $(\alpha\beta)_{ij}$ denotes the interaction effect, and ϵ_{ijk} is a random error term. Model (2.11) can be written in vector form as

$$\mathbf{y} = \mathbf{H}_0\tau_0 + \mathbf{H}_1\tau_1 + \mathbf{H}_2\boldsymbol{\tau}_2 + \mathbf{H}_3\boldsymbol{\tau}_3 + \mathbf{H}_4\boldsymbol{\tau}_4, \tag{2.12}$$

where $\mathbf{y}$ is the vector of observations, $\tau_0 = \mu$, $\tau_1 = (\alpha_1, \alpha_2, \ldots, \alpha_a)'$, $\tau_2 = (\beta_1, \beta_2, \ldots, \beta_b)'$, $\tau_3 = [(\alpha\beta)_{11}, (\alpha\beta)_{12}, \ldots, (\alpha\beta)_{ab}]'$, and $\tau_4 = (\epsilon_{111}, \epsilon_{112}, \ldots, \epsilon_{abn})'$. The matrices $\mathbf{H}_i$ $(i = 0, 1, 2, 3, 4)$ can be expressed as direct products of the form

$$\mathbf{H}_0 = \mathbf{1}_a \otimes \mathbf{1}_b \otimes \mathbf{1}_n,$$
$$\mathbf{H}_1 = \mathbf{I}_a \otimes \mathbf{1}_b \otimes \mathbf{1}_n,$$
$$\mathbf{H}_2 = \mathbf{1}_a \otimes \mathbf{I}_b \otimes \mathbf{1}_n,$$
$$\mathbf{H}_3 = \mathbf{I}_a \otimes \mathbf{I}_b \otimes \mathbf{1}_n,$$
$$\mathbf{H}_4 = \mathbf{I}_a \otimes \mathbf{I}_b \otimes \mathbf{I}_n.$$

In general, any balanced linear model can be written in vector form as

$$\mathbf{y} = \sum_{l=0}^{\nu} \mathbf{H}_l \tau_l, \tag{2.13}$$

where $\mathbf{H}_l$ $(l = 0, 1, \ldots, \nu)$ is a direct product of identity matrices and vectors of ones (see Khuri, 1982). If $\tau_0, \tau_1, \ldots, \tau_\theta$ $(\theta < \nu - 1)$ are fixed unknown parameter vectors (fixed effects), and $\tau_{\theta+1}, \tau_{\theta+2}, \ldots, \tau_\nu$ are random vectors (random effects), then model (2.11) is called a balanced mixed model. Furthermore, if we assume that the random effects are independent and have the normal distributions $N(\mathbf{0}, \sigma_l^2 \mathbf{I}_{c_l})$, where c_l is the number of columns of $\mathbf{H}_l$, $l = \theta + 1, \theta + 2, \ldots, \nu$, then, because model (2.11) is balanced, its statistical analysis becomes very simple. Here, the σ_l^2's are called the model's variance components. A balanced mixed model can be written as

$$\mathbf{y} = \mathbf{X}\mathbf{g} + \mathbf{Z}\mathbf{h} \tag{2.14}$$

where $\mathbf{X}\mathbf{g} = \sum_{l=0}^{\theta} \mathbf{H}_l \tau_l$ is the fixed portion of the model, and $\mathbf{Z}\mathbf{h} = \sum_{l=\theta+1}^{\nu} \mathbf{H}_l \tau_l$ is its random portion. The variance–covariance matrix of $\mathbf{y}$ is given by

$$\mathbf{\Sigma} = \sum_{l=\theta+1}^{\nu} \mathbf{A}_l \sigma_l^2,$$

where $\mathbf{A}_l = \mathbf{H}_l \mathbf{H}_l'$ $(l = \theta + 1, \theta + 2, \ldots, \nu)$. Note that $\mathbf{A}_l \mathbf{A}_p = \mathbf{A}_p \mathbf{A}_l$ for all $l \neq p$. Hence, the matrices $\mathbf{A}_l$ can be diagonalized simultaneously (see Theorem 2.3.15).

If $\mathbf{y}'\mathbf{A}\mathbf{y}$ is a quadratic form in $\mathbf{y}$, then $\mathbf{y}'\mathbf{A}\mathbf{y}$ is distributed as a noncentral chi-squared variate $\chi_m'^2(\eta)$ if and only if $\mathbf{A}\mathbf{\Sigma}$ is idempotent of rank m, where η is the noncentrality parameter and is given by $\eta = \mathbf{g}'\mathbf{X}'\mathbf{A}\mathbf{X}\mathbf{g}$ (see Searle, 1971, Section 2.5).

The total sum of squares, $\mathbf{y}'\mathbf{y}$, can be uniquely partitioned as

$$\mathbf{y}'\mathbf{y} = \sum_{l=0}^{\nu} \mathbf{y}'\mathbf{P}_l\mathbf{y},$$

where the $\mathbf{P}_l$'s are idempotent matrices such that $\mathbf{P}_l\mathbf{P}_s = 0$ for all $l \neq s$ (see Khuri, 1982). The quadratic form $\mathbf{y}'\mathbf{P}_l\mathbf{y}$ $(l = 0, 1, \ldots, \nu)$ is positive semidefinite and represents the sum of squares for the lth effect in model (2.13).

Theorem 2.4.1. Consider the balanced mixed model (2.14), where the random effects are assumed to be independently and normally distributed with zero means and variance–covariance matrices $\sigma_l^2\mathbf{I}_{c_l}$ $(l = \theta + 1, \theta + 2, \ldots, \nu)$. Then we have the following:

1. $\mathbf{y}'\mathbf{P}_0\mathbf{y}, \mathbf{y}'\mathbf{P}_1\mathbf{y}, \ldots, \mathbf{y}'\mathbf{P}_\nu\mathbf{y}$ are statistically independent.
2. $\mathbf{y}'\mathbf{P}_l\mathbf{y}/\delta_l$ is distributed as a noncentral chi-squared variate with degrees of freedom equal to the rank of $\mathbf{P}_l$ and noncentrality parameter given by $\eta_l = \mathbf{g}'\mathbf{X}'\mathbf{P}_l\mathbf{X}\mathbf{g}/\delta_l$ for $l = 0, 1, \ldots, \theta$, where δ_l is a particular linear combination of the variance components $\sigma_{\theta+1}^2, \sigma_{\theta+2}^2, \ldots, \sigma_\nu^2$. However, for $l = \theta + 1, \theta + 2, \ldots, \nu$, that is, for the random effects, $\mathbf{y}'\mathbf{P}_l\mathbf{y}/\delta_l$ is distributed as a central chi-squared variate with m_l degrees of freedom, where $m_l = r(\mathbf{P}_l)$.

Proof. See Theorem 4.1 in Khuri (1982). □

Theorem 2.4.1 provides the basis for a complete analysis of any balanced mixed model, as it can be used to obtain exact tests for testing the significance of the fixed effects and the variance components.

A linear function $\mathbf{a}'\mathbf{g}$, of $\mathbf{g}$ in model (2.14), is estimable if there exists a linear function, $\mathbf{c}'\mathbf{y}$, of the observations such that $E(\mathbf{c}'\mathbf{y}) = \mathbf{a}'\mathbf{g}$. In Searle (1971, Section 5.4) it is shown that $\mathbf{a}'\mathbf{g}$ is estimable if and only if $\mathbf{a}'$ belongs to the linear span of the rows of $\mathbf{X}$. In Khuri (1984) we have the following theorem:

Theorem 2.4.2. Consider the balanced mixed model in (2.14). Then we have the following:

1. $r(\mathbf{P}_l\mathbf{X}) = r(\mathbf{P}_l)$, $l = 0, 1, \ldots, \theta$.
2. $r(\mathbf{X}) = \sum_{l=0}^{\theta} r(\mathbf{P}_l\mathbf{X})$.
3. $\mathbf{P}_0\mathbf{X}\mathbf{g}, \mathbf{P}_1\mathbf{X}\mathbf{g}, \ldots, \mathbf{P}_\theta\mathbf{X}\mathbf{g}$ are linearly independent and span the space of all estimable linear functions of $\mathbf{g}$.

Theorem 2.4.2 is useful in identifying a basis of estimable linear functions of the fixed effects in model (2.14).

2.4.2. The Singular-Value Decomposition

The singular-value decomposition of a matrix is far more useful, both in statistics and in matrix algebra, then is commonly realized. For example, it

plays a significant role in regression analysis. Let us consider the linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.15}$$

where $\mathbf{y}$ is a vector of n observations, $\mathbf{X}$ is an $n \times p$ $(n \geq p)$ matrix consisting of known constants, $\boldsymbol{\beta}$ is an unknown parameter vector, and $\boldsymbol{\epsilon}$ is a random error vector. Using Theorem 2.3.11, the matrix $\mathbf{X}'$ can be expressed as

$$\mathbf{X}' = \mathbf{P}[\mathbf{D} : \mathbf{0}]\mathbf{Q}', \tag{2.16}$$

where $\mathbf{P}$ and $\mathbf{Q}$ are orthogonal matrices of orders $p \times p$ and $n \times n$, respectively, and $\mathbf{D}$ is a diagonal matrix of order $p \times p$ consisting of nonnegative diagonal elements. These are the singular values of $\mathbf{X}$ (or of $\mathbf{X}'$) and are the positive square roots of the eigenvalues of $\mathbf{X}'\mathbf{X}$. From (2.16) we get

$$\mathbf{X} = \mathbf{Q}\begin{bmatrix} \mathbf{D} \\ \mathbf{0}' \end{bmatrix}\mathbf{P}'. \tag{2.17}$$

If the columns of $\mathbf{X}$ are linearly related, then they are said to be multicollinear. In this case, $\mathbf{X}$ has rank r $(< p)$, and the columns of $\mathbf{X}$ belong to a vector subspace of dimension r. At least one of the eigenvalues of $\mathbf{X}'\mathbf{X}$, and hence at least one of the singular values of $\mathbf{X}$, will be equal to zero. In practice, such exact multicollinearities rarely occur in statistical applications. Rather, the columns of $\mathbf{X}$ may be "nearly" linearly related. In this case, the rank of $\mathbf{X}$ is p, but some of the singular values of $\mathbf{X}$ will be "near zero." We shall use the term multicollinearity in a broader sense to describe the latter situation. It is also common to use the term "ill conditioning" to refer to the same situation.

The presence of multicollinearities in $\mathbf{X}$ can have adverse effects on the least-squares estimate, $\hat{\boldsymbol{\beta}}$, of $\boldsymbol{\beta}$ in (2.15). This can be easily seen from the fact that $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ and $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\sigma^2$, where σ^2 is the error variance. Large variances associated with the elements of $\hat{\boldsymbol{\beta}}$ can therefore be expected when the columns of $\mathbf{X}$ are multicollinear. This causes $\hat{\boldsymbol{\beta}}$ to become an unreliable estimate of $\boldsymbol{\beta}$. For a detailed study of multicollinearity and its effects, see Belsley, Kuh, and Welsch (1980, Chapter 3), Montgomery and Peck (1982, Chapter 8), and Myers (1990, Chapter 3).

The singular-value decomposition of $\mathbf{X}$ can provide useful information for detecting multicollinearity, as we shall now see. Let us suppose that the columns of $\mathbf{X}$ are multicollinear. Because of this, some of the singular values of $\mathbf{X}$, say p_2 $(< p)$ of them, will be "near zero." Let us partition $\mathbf{D}$ in (2.17) as

$$\mathbf{D} = \begin{bmatrix} \mathbf{D}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_2 \end{bmatrix},$$

where $\mathbf{D}_1$ and $\mathbf{D}_2$ are of orders $p_1 \times p_1$ and $p_2 \times p_2$ ($p_1 = p - p_2$), respectively. The diagonal elements of D_2 consist of those singular values of $\mathbf{X}$ labeled as "near zero." Let us now write (2.17) as

$$\mathbf{XP} = \mathbf{Q}\begin{bmatrix} \mathbf{D}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_2 \\ \mathbf{0} & \mathbf{0} \end{bmatrix}. \tag{2.18}$$

Let us next partition $\mathbf{P}$ and $\mathbf{Q}$ as $\mathbf{P} = [\mathbf{P}_1 : \mathbf{P}_2], \mathbf{Q} = [\mathbf{Q}_1 : \mathbf{Q}_2]$, where $\mathbf{P}_1$ and $\mathbf{P}_2$ have p_1 and p_2 columns, respectively, and $\mathbf{Q}_1$ and $\mathbf{Q}_2$ have p_1 and $n - p_1$ columns, respectively. From (2.18) we conclude that

$$\mathbf{XP}_1 = \mathbf{Q}_1\mathbf{D}_1, \tag{2.19}$$

$$\mathbf{XP}_2 \approx \mathbf{0}, \tag{2.20}$$

where $\approx$ represents approximate equality. The matrix $\mathbf{XP}_2$ is "near zero" because of the smallness of the diagonal elements of $\mathbf{D}_2$.

We note from (2.20) that each column of $\mathbf{P}_2$ provides a "near"-linear relationship among the columns of $\mathbf{X}$. If (2.20) were an exact equality, then the columns of $\mathbf{P}_2$ would provide an orthonormal basis for the null space of $\mathbf{X}$.

We have mentioned that the presence of multicollinearity is indicated by the "smallness" of the singular values of $\mathbf{X}$. The problem now is to determine what "small" is. For this purpose it is common in statistics to use the condition number of $\mathbf{X}$, denoted by $\kappa(\mathbf{X})$. By definition

$$\kappa(\mathbf{X}) = \frac{\lambda_{\max}}{\lambda_{\min}},$$

where $\lambda_{\max}$ and $\lambda_{\min}$ are, respectively, the largest and smallest singular values of $\mathbf{X}$. Since the singular values of $\mathbf{X}$ are the positive square roots of the eigenvalues of $\mathbf{X}'\mathbf{X}$, then $\kappa(\mathbf{X})$ can also be written as

$$\kappa(\mathbf{X}) = \sqrt{\frac{e_{\max}(\mathbf{X}'\mathbf{X})}{e_{\min}(\mathbf{X}'\mathbf{X})}}.$$

If $\kappa(\mathbf{X})$ is less than 10, then there is no serious problem with multicollinearity. Values of $\kappa(\mathbf{X})$ between 10 and 30 indicate moderate to strong multicollinearity, and if $\kappa > 30$, severe multicollinearity is implied.

More detailed discussions concerning the use of the singular-value decomposition in regression can be found in Mandel (1982). See also Lowerre (1982). Good (1969) described several applications of this decomposition in statistics and in matrix algebra.

2.4.3. Extrema of Quadratic Forms

In many statistical problems there is a need to find the extremum (maximum or minimum) of a quadratic form or a ratio of quadratic forms. Let us, for example, consider the following problem:

Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be a collection of random vectors, all having the same number of elements. Suppose that these vectors are independently and identically distributed (i.i.d.) as $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown. Consider testing the hypothesis H_0: $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ versus its alternative H_a: $\boldsymbol{\mu} \neq \boldsymbol{\mu}_0$, where $\boldsymbol{\mu}_0$ is some hypothesized value of $\boldsymbol{\mu}$. We need to develop a test statistic for testing H_0.

The multivariate hypothesis H_0 is true if and only if the univariate hypotheses

$$H_0(\lambda)\colon \boldsymbol{\lambda}'\boldsymbol{\mu} = \boldsymbol{\lambda}'\boldsymbol{\mu}_0$$

are true for all $\boldsymbol{\lambda} \neq \mathbf{0}$. A test statistic for testing $H_0(\boldsymbol{\lambda})$ is the following:

$$t(\boldsymbol{\lambda}) = \frac{\boldsymbol{\lambda}'(\overline{\mathbf{X}} - \boldsymbol{\mu}_0)\sqrt{n}}{\sqrt{\boldsymbol{\lambda}'\mathbf{S}\boldsymbol{\lambda}}},$$

where $\overline{\mathbf{X}} = \sum_{i=1}^{n} \mathbf{X}_i/n$ and $\mathbf{S}$ is the sample variance–covariance matrix, which is an unbiased estimator of $\boldsymbol{\Sigma}$, and is given by

$$\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})'.$$

Large values of $t^2(\boldsymbol{\lambda})$ indicate falsehood of $H_0(\boldsymbol{\lambda})$. Since H_0 is rejected if and only if $H_0(\boldsymbol{\lambda})$ is rejected for at least one $\boldsymbol{\lambda}$, then the condition to reject H_0 at the α-level is $\sup_{\boldsymbol{\lambda} \neq \mathbf{0}}[t^2(\boldsymbol{\lambda})] > c_\alpha$, where c_α is the upper $100\alpha\%$ point of the distribution of $\sup_{\boldsymbol{\lambda} \neq \mathbf{0}}[t^2(\boldsymbol{\lambda})]$. But

$$\begin{aligned}
\sup_{\boldsymbol{\lambda} \neq \mathbf{0}} [t^2(\boldsymbol{\lambda})] &= \sup_{\boldsymbol{\lambda} \neq \mathbf{0}} \frac{n|\boldsymbol{\lambda}'(\overline{\mathbf{X}} - \boldsymbol{\mu}_0)|^2}{\boldsymbol{\lambda}'\mathbf{S}\boldsymbol{\lambda}} \\
&= n \sup_{\boldsymbol{\lambda} \neq \mathbf{0}} \frac{\boldsymbol{\lambda}'(\overline{\mathbf{X}} - \boldsymbol{\mu}_0)(\overline{\mathbf{X}} - \boldsymbol{\mu}_0)'\boldsymbol{\lambda}}{\boldsymbol{\lambda}'\mathbf{S}\boldsymbol{\lambda}} \\
&= n\, e_{\max}\left[\mathbf{S}^{-1}(\overline{\mathbf{X}} - \boldsymbol{\mu}_0)(\overline{\mathbf{X}} - \boldsymbol{\mu}_0)'\right],
\end{aligned}$$

by Theorem 2.3.17.

Now,

$$\begin{aligned}
e_{\max}\left[\mathbf{S}^{-1}(\overline{\mathbf{X}} - \boldsymbol{\mu}_0)(\overline{\mathbf{X}} - \boldsymbol{\mu}_0)'\right] &= e_{\max}\left[(\overline{\mathbf{X}} - \boldsymbol{\mu}_0)'\mathbf{S}^{-1}(\overline{\mathbf{X}} - \boldsymbol{\mu}_0)\right], \\
&= (\overline{\mathbf{X}} - \boldsymbol{\mu}_0)'\mathbf{S}^{-1}(\overline{\mathbf{X}} - \boldsymbol{\mu}_0).
\end{aligned}$$

by Theorem 2.3.9.