Series/Number 07-159

A MATHEMATICAL PRIMER FOR SOCIAL STATISTICS

John Fox
McMaster University

SAGE
Los Angeles • London • New Delhi • Singapore

CONTENTS

ABOUT THE AUTHOR

John Fox is Professor of Sociology at McMaster University in Hamilton, Ontario, Canada. He was previously Professor of Sociology and of Mathematics and Statistics at York University in Toronto, where he also directed the Statistical Consulting Service at the Institute for Social Research. Professor Fox earned a Ph.D. in Sociology from the University of Michigan in 1972. He has delivered numerous lectures and workshops on statistical topics, at such places as the summer program of the Inter-University Consortium for Political and Social Research and the annual meetings of the American Sociological Association. His recent and current work includes research on statistical methods (for example, work on three-dimensional statistical graphs) and on Canadian society (for example, a study of political polls in the 1995 Quebec sovereignty referendum). He is author of many articles, in such journals as *Sociological Methodology*, *The Journal of Computational and Graphical Statistics*, *The Journal of the American Statistical Association*, *The Canadian Review of Sociology and Anthropology*, and *The Canadian Journal of Sociology*. He has written several other books, including *Applied Regression Analysis and Generalized Linear Models, Second Edition* (Sage, 2008), *Nonparametric Simple Regression* (Sage, 2000), and *Multiple and Generalized Nonparametric Regression* (Sage, 2000).

SERIES EDITOR'S INTRODUCTION

A fellow graduate student in the same PhD sociology program once explained to me, having decided to take a foundation course in the statistics department, "Because there's always this wall that I've run against when trying to learn advanced quantitative methods." A course offered by the statistics department has a stronger mathematical foundation than a similar one offered by a social science program, thus enabling my friend to overcome the barrier in his background knowledge in statistics. Granted, the *Quantitative Applications in the Social Sciences* series is designed for readers who do not have very extensive mathematical or statistical training. Still, many of our recent titles including those on robust regression and latent growth curve models are at a level high enough to leave some gaps in a common reader's knowledge when reading the books.

John Fox's *A Mathematical Primer for Social Statistics* is written with people who may have a wall in their learning path or some knowledge gaps in mind. The volume covers many often ignored yet important topics (such as the topics of matrix and linear algebra, of calculus, and of probability theory and certain statistical distributions) in both the subject areas of mathematics and statistics, those with frequent appearances in many statistics books and articles that are assumed that the reader has learned in a previous life. For most social science readers, the assumption is simply not true.

When I took up the editorship in early 2004, I planned to publish in the series three types of books, those that will continue to follow or even push the frontier, those that will fill the gaps in currently well-known methodology, and those that will get back to basics. I was very pleased when I learned about John's project, I quickly encouraged him to complete it as I knew the book would not only be an important addition to the series but would also help fulfill my vision as editor, notably the plan to get back to basics. Indeed, both reviewers of the manuscript wished that such a book had been available years before, either when they were learning statistical topics in their own education or when they were preparing their lecture notes and handouts for teaching their first quantitative methods courses.

As the reviewers summed up, "This book will prove to be extremely useful for graduate students and practitioners of social statistics," and "will offer a highly-welcome and valuable addition" to the series.

—Tim Futing Liao
Series Editor

PREFACE

Statistics is not mathematics. Math is central to the development, communication, and understanding of statistics, but applied statistics—the kind of statistics of most interest to social scientists—is not about proving abstract theorems, but about analyzing data.

Typical introductory statistics courses taught to social science students use only very basic mathematics—arithmetic, simple formulas, the ability to interpret a graph. There are good reasons for this: Most social science students have weak backgrounds in mathematics. Even more important, however, the fundamental goals of a basic statistics course (or at least what in my opinion should be the fundamental goals) are to convey the role of statistical methods in gathering and summarizing data, and the essential ideas of statistical inference. Accomplishing these goals is sufficiently challenging without drowning the big ideas in a sea of equations. I believe, incidentally, that this is the case even for students who have strong foundations in mathematics.

Once beyond the introductory level, and perhaps a second course in applied regression analysis, the situation changes: Insufficient grounding in mathematics makes it difficult to proceed in applied statistics. The good news, however, is that a relatively modest background in intermediate-level mathematics suffices for the study of a great deal of statistics. Often, all that is needed is an understanding of basic mathematical ideas, familiarity with some important facts, and an ability to manipulate simple equations. This book aims to provide that basic background.

The book originated in online appendices that I wrote for the second edition of my applied regression text (Fox, 2008). I felt initially that some readers might prefer a printed and bound copy of the appendices to downloading them from the Internet. It then occurred to me that the appendices might prove more generally useful, and ultimately I augmented them with material that was not directly relevant to my applied regression book but which is important to other statistical methods that are widely employed in the social sciences. This book, therefore, includes material not in the original appendices.

The book covers three areas of mathematics that are of central importance to applied statistics:

- Chapter 1 takes up matrices, linear algebra, and vector geometry. Matrices, which are rectangular arrays of numbers, are a natural representation of most

statistical data, and consequently the arithmetic and algebra of matrices is the natural language for developing most statistical methods. Beyond the basic level, matrices are omnipresent in statistics, and therefore some acquaintance with matrices is necessary for reading statistical material. The closely related areas of linear algebra and its visual representation, vector geometry, are also central to the development and understanding of many statistical methods.

- Chapter 2 introduces the basic ideas of differential and integral calculus. Here, the emphasis is on fundamental concepts and simple methods. Differential calculus is frequently used in statistics for optimization problems—that is, minimization and maximization: Think, for example, of the method of *least* squares or of *maximum*-likelihood estimation. Integral calculus figures prominently in probability theory, which is fundamentally tied to statistical modeling and statistical inference. Although the presentation of calculus in this book is elementary, I do cover topics important to statistics, such as multivariable and matrix calculus, that, while not fundamentally difficult, are often deferred to advanced treatments of the subject.
- Chapter 3 develops probability theory and the theory of statistical estimation, including a description of important probability distributions, the central method of maximum likelihood, and the basics of Bayesian statistical inference. The ideas in this chapter feature prominently in applied statistics, and indeed the chapter represents a kind of "crash course" in some of the fundamentals of mathematical statistics.

The fourth and final chapter illustrates the use of the mathematics for applied statistics by briefly developing the seminal statistical method of linear least-squares regression and deriving some of its properties. The object of this chapter is to show how the math can be put to work.

It is, all told, remarkable how far one can get in applied statistics with a modicum of mathematics—the modicum that this book supplies. This is the book that I wish I had when I started to study statistics seriously. I hope that it will prove helpful to you, both on initial reading and as a reference.

Notation

Specific notation is introduced at various points in the text. Throughout the text, I adhere to the following general conventions, with few exceptions. [Examples are shown in brackets.]

- Known scalar constants (including subscripts) are represented by lowercase italic letters [a, b, x_i].
- Observable scalar random variables are represented by uppercase italic letters [X, Y_i]. Where it is necessary to make the distinction, *specific values* of random variables are represented as constants [x, y_i].

TABLE 1
The Greek Alphabet With Roman "Equivalents"

Greek Letter			*Roman Equivalent*	
Lowercase	*Uppercase*		*Phonetic*	*Other*
α	A	alpha	a	
β	B	beta	b	
γ	Γ	gamma	g, n	c
δ	Δ	delta	d	
ε	E	epsilon	e	
ζ	Z	zeta	z	
η	H	eta	e	
θ	Θ	theta	th	
ι	I	iota	i	
κ	K	kappa	k	
λ	Λ	lambda	l	
μ	M	mu	m	
ν	N	nu	n	
ξ	Ξ	xi	x	
o	O	omicron	o	
π	Π	pi	p	
ρ	P	rho	r	
σ	Σ	sigma	s	
τ	T	tau	t	
υ	Υ	upsilon	y, u	
ϕ	Φ	phi	ph	
χ	X	chi	ch	x
ψ	Ψ	psi	ps	
ω	Ω	omega	o	w

- Scalar parameters are represented by lowercase Greek letters [α, β, γ_2]. (See the Greek alphabet in Table 1.) Their estimators are generally denoted by "corresponding" italic characters [A, B, C_2], or by Greek letters with "hats" [$\widehat{\alpha}$, $\widehat{\beta}$, $\widehat{\gamma}_2$].
- Unobservable scalar random variables are also represented by lowercase Greek letters [ε_i].
- Vectors and matrices are represented by boldface characters—lowercase for vectors [$\mathbf{x}_1$, $\boldsymbol{\beta}$], uppercase for matrices [$\mathbf{X}$, $\boldsymbol{\Sigma}$]. In a statistical context, Roman letters are used for constants and observable random variables [$\mathbf{y}$, $\mathbf{x}_1$, $\mathbf{X}$], and Greek letters are used for parameters and unobservable random variables [$\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$, $\boldsymbol{\varepsilon}$]. It is occasionally convenient to show the order of a vector or matrix below the matrix [$\underset{(n\times 1)}{\boldsymbol{\varepsilon}}$, $\underset{(n\times k+1)}{\mathbf{X}}$]. The order of an identity matrix is given

by a subscript [$\mathbf{I}_n$]. A zero matrix or vector is represented by a boldface 0 [$\mathbf{0}$]; a vector of 1s is represented by a boldface $\mathbf{1}$, possibly subscripted with its number of elements [$\mathbf{1}_n$]. Vectors are column vectors, unless they are explicitly transposed [column: $\mathbf{x}$; row: $\mathbf{x}'$].

- The symbol $\equiv$ can be read as "is defined by," or "is equal to by definition" [$\overline{X} \equiv (\sum X_i)/n$].
- The symbol $\approx$ means "is approximately equal to" [$\pi \approx 3.14159$].
- The symbol $\propto$ means "is proportional to" [$p(\alpha|D) \propto L(\alpha)p(\alpha)$].
- The symbol $\sim$ means "is distributed as" [$\varepsilon_i \sim N(0, \sigma^2)$].
- The operator $E(\,)$ denotes the expectation of a scalar, vector, or matrix random variable [$E(Y_i)$, $E(\boldsymbol{\varepsilon})$, $E(\mathbf{X})$].
- The operator $V(\,)$ denotes the variance of a scalar random variable or the variance-covariance matrix of a vector random variable [$V(\varepsilon_i)$, $V(\mathbf{b})$].
- Estimated variances or variance-covariance matrices are indicated by a circumflex ("hat") placed over the variance operator [$\widehat{V}(\varepsilon_i)$, $\widehat{V}(\mathbf{b})$].
- The operator $C(\,)$ gives the covariance of two scalar random variables or the covariance matrix of two vector random variables [$C(X, Y)$, $C(\mathbf{x}, \mathbf{y})$].
- The operators $\mathcal{E}(\,)$ and $\mathcal{V}(\,)$ denote asymptotic expectation and variance, respectively. Their usage is similar to that of $E(\,)$ and $V(\,)$ [$\mathcal{E}(B)$, $\mathcal{V}(\widehat{\boldsymbol{\beta}})$, $\widehat{\mathcal{V}}(B)$].
- Probability limits are specified by plim [plim $b = \beta$].
- Standard mathematical functions are shown in lowercase [cos W, trace($\mathbf{A}$)]. The base of the log function is always specified explicitly, unless it is irrelevant [$\log_e L$, $\log_{10} X$]. The exponential function $\exp(x)$ represents e^x.
- The summation sign $\sum$ is used to denote continued addition [$\sum_{i=1}^{n} X_i \equiv X_1 + X_2 + \cdots + X_n$]. Often, the range of the index is suppressed if it is clear from the context [$\sum_i X_i$], and the index may be suppressed as well [$\sum X_i$]. The symbol $\prod$ similarly indicates continued multiplication [$\prod_{i=1}^{n} p(Y_i) \equiv p(Y_1) \times p(Y_2) \times \cdots \times p(Y_n)$].
- The symbol ∂ denotes the partial derivative [$\partial f(x_1, x_2)/\partial x_1$].
- To avoid awkward and repetitive phrasing in the statement of definitions and results, the words "if" and "when" are understood to mean "if and only if," unless explicitly indicated to the contrary. Terms are generally set in *italics* when they are introduced. ["Two vectors are *orthogonal* if their inner product is 0."]

Acknowledgments

I am grateful to Robert Andersen of the University of Toronto and two anonymous reviewers for helpful comments on a draft of this book, and to Tim Liao, the editor of the QASS series, and Vicki Knight of Sage Publications for their help and support. I would also like to acknowledge the support of the Social Sciences and Humanities Research Council of Canada.

A MATHEMATICAL PRIMER FOR SOCIAL STATISTICS

JOHN FOX
McMaster University

CHAPTER 1. MATRICES, LINEAR ALGEBRA, AND VECTOR GEOMETRY

Matrices provide a natural notation for much of statistics; the algebra of linear statistical models is linear algebra; and vector geometry is a powerful conceptual tool for understanding linear algebra and for visualizing many aspects of linear models. The purpose of this chapter is to present basic concepts and results concerning matrices, linear algebra, and vector geometry. The focus is on topics that are employed widely in social statistics, and the style of presentation is informal rather than mathematically rigorous: At points, results are stated without proof; at other points, proofs are outlined; often, results are justified intuitively. Readers interested in pursuing linear algebra at greater depth might profitably make reference to one of the many available texts on the subject, each of which develops in greater detail most of the topics presented here (see, e.g., the recommended readings at the end of the chapter).

The first section of the chapter develops elementary matrix algebra. Sections 1.2 and 1.3 introduce vector geometry and vector spaces. Section 1.4 discusses the related topics of matrix rank and the solution of linear simultaneous equations. Sections 1.5 and 1.6 deal with eigenvalues, eigenvectors, quadratic forms, and positive-definite matrices.

1.1 Matrices

1.1.1 Introducing the Actors: Basic Definitions

A *matrix* is a rectangular table of numbers or of numerical variables; for example,

$$\underset{(4\times 3)}{\mathbf{X}} = \begin{bmatrix} 1 & -2 & 3 \\ 4 & -5 & -6 \\ 7 & 8 & 9 \\ 0 & 0 & 10 \end{bmatrix} \tag{1.1}$$

or, more generally,

$$\underset{(m\times n)}{\mathbf{A}} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \tag{1.2}$$

A matrix such as this with m rows and n columns is said to be of *order* m by n, written $(m \times n)$. For clarity, I at times indicate the order of a matrix below the matrix, as in Equations 1.1 and 1.2. Each entry or element of a matrix may be subscripted by its row and column indices: a_{ij} is the entry in the ith row and jth column of the matrix $\mathbf{A}$. Individual numbers, such as the entries of a matrix, are termed *scalars*. Sometimes, for compactness, I specify a matrix by enclosing its typical element in braces; for example, $\underset{(m\times n)}{\mathbf{A}} = \{a_{ij}\}$ is equivalent to Equation 1.2.

A matrix consisting of one column is called a *column vector*; for example,

$$\underset{(m\times 1)}{\mathbf{a}} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix}$$

Likewise, a matrix consisting of one row is called a *row vector*,

$$\mathbf{b}' = [b_1, b_2, \cdots, b_n]$$

In specifying a row vector, I typically place commas between its elements for clarity.

The *transpose* of a matrix $\mathbf{A}$, denoted $\mathbf{A}'$, is formed from $\mathbf{A}$ so that the ith *row* of $\mathbf{A}'$ consists of the elements of the ith *column* of $\mathbf{A}$; thus (using the matrices in Equations 1.1 and 1.2),

$$\underset{(3\times 4)}{\mathbf{X}'} = \begin{bmatrix} 1 & 4 & 7 & 0 \\ -2 & -5 & 8 & 0 \\ 3 & -6 & 9 & 10 \end{bmatrix}$$

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

Note that $(\mathbf{A}')' = \mathbf{A}$. I adopt the convention that a vector is a column vector (such as $\mathbf{a}$ above) unless it is explicitly transposed (such as $\mathbf{b}'$).

A *square matrix of order n*, as the name implies, has n rows and n columns. The entries a_{ii} (i.e., $a_{11}, a_{22}, \ldots, a_{nn}$) of a square matrix $\mathbf{A}$ comprise the *main diagonal* of the matrix. The sum of the diagonal elements is the *trace* of the matrix:

$$\text{trace}(A) \equiv \sum_{i=1}^{n} a_{ii}$$

For example, the square matrix

$$\underset{(3\times 3)}{\mathbf{B}} = \begin{bmatrix} -5 & 1 & 3 \\ 2 & 2 & 6 \\ 7 & 3 & -4 \end{bmatrix}$$

has diagonal elements, $-5, 2$, and -4, and $\text{trace}(\mathbf{B}) = \sum_{i=1}^{3} b_{ii} = -5 + 2 - 4 = -7$.

A square matrix $\mathbf{A}$ is *symmetric* if $\mathbf{A} = \mathbf{A}'$, that is, when $a_{ij} = a_{ji}$ for all i and j. Consequently, the matrix $\mathbf{B}$ (above) is not symmetric, while the matrix

$$\mathbf{C} = \begin{bmatrix} -5 & 1 & 3 \\ 1 & 2 & 6 \\ 3 & 6 & -4 \end{bmatrix}$$

is symmetric. Many matrices that appear in statistical applications are symmetric—for example, correlation matrices, covariance matrices, and matrices of sums of squares and cross-products.

An *upper-triangular matrix* is a square matrix with zeroes below its main diagonal:

$$\underset{(n\times n)}{\mathbf{U}} = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ 0 & u_{22} & \cdots & u_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & u_{nn} \end{bmatrix}$$

Similarly, a *lower-triangular matrix* is a square matrix of the form

$$\underset{(n\times n)}{\mathbf{L}} = \begin{bmatrix} l_{11} & 0 & \cdots & 0 \\ l_{21} & l_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & \cdots & l_{nn} \end{bmatrix}$$

A square matrix is *diagonal* if all entries except those on its main diagonal are zero; thus,

$$\underset{(n\times n)}{\mathbf{D}} = \begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_n \end{bmatrix}$$

For compactness, I may write $\mathbf{D} = \text{diag}(d_1, d_2, \ldots, d_n)$. A *scalar matrix* is a diagonal matrix all of whose diagonal entries are equal: $\mathbf{S} = \text{diag}(s, s, \ldots, s)$. An especially important family of scalar matrices are the *identity matrices* $\mathbf{I}$, which have ones on the main diagonal:

$$\underset{(n\times n)}{\mathbf{I}} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

I write $\mathbf{I}_n$ for $\underset{(n\times n)}{\mathbf{I}}$.

Two other special matrices are the family of *zero matrices* $\mathbf{0}$, all of whose entries are zero, and the vectors $\mathbf{1}$, all of whose entries are one. I write $\mathbf{1}_n$ for the vector of ones with n entries; for example $\mathbf{1}_4 = [1, 1, 1, 1]'$. Although the identity matrices, the zero matrices, and the vectors $\mathbf{1}$ are *families* of

matrices, it is often convenient to refer to the matrices in the singular, for example, to *the* identity matrix.

A *partitioned matrix* is a matrix whose elements are organized into *submatrices*; for example,

$$\underset{(4\times 3)}{\mathbf{A}} = \left[\begin{array}{cc|c} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \\ \hline a_{41} & a_{42} & a_{43} \end{array}\right] = \left[\begin{array}{c|c} \underset{(3\times 2)}{\mathbf{A}_{11}} & \underset{(3\times 1)}{\mathbf{A}_{12}} \\ \hline \underset{(1\times 2)}{\mathbf{A}_{21}} & \underset{(1\times 1)}{\mathbf{A}_{22}} \end{array}\right]$$

where the submatrix

$$\mathbf{A}_{11} \equiv \left[\begin{array}{cc} a_{11} & a_{21} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{array}\right]$$

and $\mathbf{A}_{12}$, $\mathbf{A}_{21}$, and $\mathbf{A}_{22}$ are similarly defined. When there is no possibility of confusion, I omit the lines separating the submatrices. If a matrix is partitioned vertically but not horizontally, then I separate its submatrices by commas; for example, $\underset{(m\times n+p)}{\mathbf{C}} = \left[\underset{(m\times n)}{\mathbf{C}_1}, \underset{(m\times p)}{\mathbf{C}_2}\right]$.

1.1.2 Simple Matrix Arithmetic

Two matrices are equal if they are of the same order and all corresponding entries are equal (a definition used implicitly in Section 1.1.1).

Two matrices may be added only if they are of the same order; then their sum is formed by adding corresponding elements. Thus, if $\mathbf{A}$ and $\mathbf{B}$ are of order $(m \times n)$, then $\mathbf{C} = \mathbf{A} + \mathbf{B}$ is also of order $(m \times n)$, with $c_{ij} = a_{ij} + b_{ij}$. Likewise, if $\mathbf{D} = \mathbf{A} - \mathbf{B}$, then $\mathbf{D}$ is of order $(m \times n)$, with $d_{ij} = a_{ij} - b_{ij}$. The negative of a matrix $\mathbf{A}$, that is, $\mathbf{E} = -\mathbf{A}$, is of the same order as $\mathbf{A}$, with elements $e_{ij} = -a_{ij}$. For example, for matrices

$$\underset{(2\times 3)}{\mathbf{A}} = \left[\begin{array}{ccc} 1 & 2 & 3 \\ 4 & 5 & 6 \end{array}\right]$$

and

$$\underset{(2\times 3)}{\mathbf{B}} = \left[\begin{array}{ccc} -5 & 1 & 2 \\ 3 & 0 & -4 \end{array}\right]$$

we have

$$\underset{(2\times3)}{\mathbf{C}} = \mathbf{A} + \mathbf{B} = \begin{bmatrix} -4 & 3 & 5 \\ 7 & 5 & 2 \end{bmatrix}$$

$$\underset{(2\times3)}{\mathbf{D}} = \mathbf{A} - \mathbf{B} = \begin{bmatrix} 6 & 1 & 1 \\ 1 & 5 & 10 \end{bmatrix}$$

$$\underset{(2\times3)}{\mathbf{E}} = -\mathbf{B} = \begin{bmatrix} 5 & -1 & -2 \\ -3 & 0 & 4 \end{bmatrix}$$

Because they are element-wise operations, matrix addition, subtraction, and negation follow essentially the same rules as the corresponding scalar operations; in particular,

$$\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A} \text{ (matrix addition is commutative)}$$
$$\mathbf{A} + (\mathbf{B} + \mathbf{C}) = (\mathbf{A} + \mathbf{B}) + \mathbf{C} \text{ (matrix addition is associative)}$$
$$\mathbf{A} - \mathbf{B} = \mathbf{A} + (-\mathbf{B}) = -(\mathbf{B} - \mathbf{A})$$
$$\mathbf{A} - \mathbf{A} = \mathbf{0}$$
$$\mathbf{A} + \mathbf{0} = \mathbf{A}$$
$$-(-\mathbf{A}) = \mathbf{A}$$
$$(\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}'$$

The product of a scalar c and an $(m \times n)$ matrix $\mathbf{A}$ is an $(m \times n)$ matrix $\mathbf{B} = c\mathbf{A}$ in which $b_{ij} = ca_{ij}$. Continuing the preceding examples:

$$\underset{(2\times3)}{\mathbf{F}} = 3 \times \mathbf{B} = \mathbf{B} \times 3 = \begin{bmatrix} -15 & 3 & 6 \\ 9 & 0 & -12 \end{bmatrix}$$

The product of a scalar and a matrix obeys the following rules:

$$c\mathbf{A} = \mathbf{A}c \text{ (commutative)}$$
$$\mathbf{A}(b + c) = \mathbf{A}b + \mathbf{A}c \text{ (distributes over scalar addition)}$$
$$c(\mathbf{A} + \mathbf{B}) = c\mathbf{A} + c\mathbf{B} \text{ (distributes over matrix addition)}$$
$$0\mathbf{A} = \mathbf{0}$$
$$1\mathbf{A} = \mathbf{A}$$
$$(-1)\mathbf{A} = -\mathbf{A}$$

where, note, b, c, 0, 1, and -1 are scalars, and $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{0}$ are matrices of the same order.

The *inner product* (or *dot product*) of two vectors (each with n entries), say $\underset{(1\times n)}{\mathbf{a}'}$ and $\underset{(n\times 1)}{\mathbf{b}}$, denoted $\mathbf{a}'\cdot\mathbf{b}$, is a scalar formed by multiplying corresponding entries of the vectors and summing the resulting products:

$$\mathbf{a}' \cdot \mathbf{b} = \sum_{i=1}^{n} a_i b_i$$

For example,

$$[2, 0, 1, 3] \cdot \begin{bmatrix} -1 \\ 6 \\ 0 \\ 9 \end{bmatrix} = 2(-1) + 0(6) + 1(0) + 3(9) = 25$$

Although this example is for the inner product of a row vector with a column vector, both vectors may be row vectors or both column vectors, as long as the two vectors have the same number of elements.

Two matrices $\mathbf{A}$ and $\mathbf{B}$ are *conformable for multiplication* in the order given (i.e., $\mathbf{AB}$) if the number of *columns* of the left-hand factor ($\mathbf{A}$) is equal to the number of *rows* of the right-hand factor ($\mathbf{B}$). Thus $\mathbf{A}$ and $\mathbf{B}$ are conformable for multiplication if $\mathbf{A}$ is of order $(m \times n)$ and $\mathbf{B}$ is of order $(n \times p)$, where m and p are unconstrained. For example,

$$\underset{(2\times 3)}{\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}} \underset{(3\times 3)}{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}}$$

are conformable for multiplication but

$$\underset{(3\times 3)}{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}} \underset{(2\times 3)}{\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}}$$

are not.

Let $\mathbf{C} = \mathbf{AB}$ be the matrix product; and let $\mathbf{a}_i'$ represent the ith *row* of $\mathbf{A}$ and $\mathbf{b}_j$ represent the jth *column* of $\mathbf{B}$. Then $\mathbf{C}$ is a matrix of order $(m \times p)$ in which

$$c_{ij} = \mathbf{a}_i' \cdot \mathbf{b}_j = \sum_{k=1}^{n} a_{ik} b_{kj}$$

Some examples:

$$\underset{(2\times 3)}{\begin{bmatrix} & \Longrightarrow & \\ 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}} \underset{(3\times 3)}{\begin{bmatrix} & 1 & 0 & 0 \\ \Downarrow & 0 & 1 & 0 \\ & 0 & 0 & 1 \end{bmatrix}}$$

$$= \underset{(2\times 3)}{\begin{bmatrix} 1(1)+2(0)+3(0), & 1(0)+2(1)+3(0), & 1(0)+2(0)+3(1) \\ 4(1)+5(0)+6(0), & 4(0)+5(1)+6(0), & 4(0)+5(0)+6(1) \end{bmatrix}}$$

$$= \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$$

$$\underset{(1\times 4)}{[\beta_0, \beta_1, \beta_2, \beta_3]} \underset{(4\times 1)}{\begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_3 \end{bmatrix}} = \underset{(1\times 1)}{[\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3]}$$

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 0 & 3 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} 4 & 5 \\ 8 & 13 \end{bmatrix} \tag{1.3}$$

$$\begin{bmatrix} 0 & 3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 9 & 12 \\ 5 & 8 \end{bmatrix}$$

$$\begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{3} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{1.4}$$

$$\begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{3} \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

In the first of these examples, the arrows indicate how the rows of the left-hand factor are multiplied into the columns of the right-hand factor.

Matrix multiplication is associative, $\mathbf{A(BC)} = \mathbf{(AB)C}$, and distributive with respect to addition:

$$\mathbf{(A+B)C} = \mathbf{AC} + \mathbf{BC}$$
$$\mathbf{A(B+C)} = \mathbf{AB} + \mathbf{AC}$$

but it is not in general commutative: If $\mathbf{A}$ is $(m \times n)$ and $\mathbf{B}$ is $(n \times p)$, then the product $\mathbf{AB}$ is defined but $\mathbf{BA}$ is defined only if $m = p$. Even so, $\mathbf{AB}$ and $\mathbf{BA}$ are of different orders (and hence are not candidates for equality) unless $m = p$. And even if $\mathbf{A}$ and $\mathbf{B}$ are square, $\mathbf{AB}$ and $\mathbf{BA}$, though of the same order, are not necessarily equal (as illustrated in Equation 1.3). If it is the case that $\mathbf{AB} = \mathbf{BA}$ (as in Equation 1.4), then the matrices $\mathbf{A}$ and $\mathbf{B}$ are said to *commute* with one another. A scalar factor, however, may be moved anywhere within a matrix product: $c\mathbf{AB} = \mathbf{A}c\mathbf{B} = \mathbf{AB}c$.

3(1)
6(1)

The identity and zero matrices play roles with respect to matrix multiplication analogous to those of the numbers 0 and 1 in scalar algebra:

$$\underset{(m\times n)}{\mathbf{A}}\ \underset{(n\times n)}{\mathbf{I}_n} = \underset{(m\times m)}{\mathbf{I}_m}\ \underset{(m\times n)}{\mathbf{A}} = \mathbf{A}$$

$$\underset{(m\times n)(n\times p)}{\mathbf{A}\ \mathbf{0}} = \underset{(m\times p)}{\mathbf{0}}$$

$$\underset{(q\times m)(m\times n)}{\mathbf{0}\ \mathbf{A}} = \underset{(q\times n)}{\mathbf{0}}$$

A further property of matrix multiplication, which has no analog in scalar algebra, is that $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$—the transpose of a product is the product of the transposes taken in the opposite order, a rule that extends to several (conformable) matrices:

(1.3)
$$(\mathbf{AB}\cdots\mathbf{F})' = \mathbf{F}'\cdots\mathbf{B}'\mathbf{A}'$$

The *powers* of a square matrix are the products of the matrix with itself. That is, $\mathbf{A}^2 = \mathbf{AA}$, $\mathbf{A}^3 = \mathbf{AAA} = \mathbf{AA}^2 = \mathbf{A}^2\mathbf{A}$, and so on. If $\mathbf{B}^2 = \mathbf{A}$, then we call $\mathbf{B}$ a *square root* of $\mathbf{A}$, which we may write as $\mathbf{A}^{1/2}$. Unlike in scalar algebra, however, the square root of a matrix is not generally unique.

(1.4)
Of course, even the scalar square root is unique only up to a change in sign. (For another kind of matrix square root, see the discussion of the Cholesky decomposition in Section 1.6.1.) If $\mathbf{A}^2 = \mathbf{A}$, then $\mathbf{A}$ is said to be *idempotent*. As in scalar algebra, and by convention, $\mathbf{A}^0 = \mathbf{I}$ (where $\mathbf{I}$ is of the same order as $\mathbf{A}$). The matrix inverse $\mathbf{A}^{-1}$ is discussed in Section 1.1.3, and is *not* $\{1/a_{ij}\}$.

e left-

For purposes of matrix addition, subtraction, and multiplication, the submatrices of partitioned matrices may be treated as if they were elements, as long as the factors are partitioned conformably. For example, if

butive

$$\mathbf{A} = \left[\begin{array}{ccc|cc} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ \hline a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \end{array}\right] = \left[\begin{array}{cc} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{array}\right]$$

and

$$\mathbf{B} = \left[\begin{array}{ccc|cc} b_{11} & b_{12} & b_{13} & b_{14} & b_{15} \\ b_{21} & b_{22} & b_{23} & b_{24} & b_{25} \\ \hline b_{31} & b_{32} & b_{33} & b_{34} & b_{35} \end{array}\right] = \left[\begin{array}{cc} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{array}\right]$$

then

$$\mathbf{A} + \mathbf{B} = \left[\begin{array}{c|c} \mathbf{A}_{11} + \mathbf{B}_{11} & \mathbf{A}_{12} + \mathbf{B}_{12} \\ \hline \mathbf{A}_{21} + \mathbf{B}_{21} & \mathbf{A}_{22} + \mathbf{B}_{22} \end{array}\right]$$

Similarly, if

$$\underset{(m+n\times p+q)}{\mathbf{A}} = \left[\begin{array}{cc} \underset{(m\times p)}{\mathbf{A}_{11}} & \underset{(m\times q)}{\mathbf{A}_{12}} \\ \underset{(n\times p)}{\mathbf{A}_{21}} & \underset{(n\times q)}{\mathbf{A}_{22}} \end{array}\right]$$

and

$$\underset{(p+q\times r+s)}{\mathbf{B}} = \left[\begin{array}{cc} \underset{(p\times r)}{\mathbf{B}_{11}} & \underset{(p\times s)}{\mathbf{B}_{12}} \\ \underset{(q\times r)}{\mathbf{B}_{21}} & \underset{(q\times s)}{\mathbf{B}_{22}} \end{array}\right]$$

then

$$\underset{(m+n\times r+s)}{\mathbf{AB}} = \left[\begin{array}{c|c} \mathbf{A}_{11}\mathbf{B}_{11} + \mathbf{A}_{12}\mathbf{B}_{21} & \mathbf{A}_{11}\mathbf{B}_{12} + \mathbf{A}_{12}\mathbf{B}_{22} \\ \hline \mathbf{A}_{21}\mathbf{B}_{11} + \mathbf{A}_{22}\mathbf{B}_{21} & \mathbf{A}_{21}\mathbf{B}_{12} + \mathbf{A}_{22}\mathbf{B}_{22} \end{array}\right]$$

The Sense Behind Matrix Multiplication The definition of matrix multiplication makes it simple to formulate systems of scalar equations as a single matrix equation, often providing a useful level of abstraction. For example, consider the following system of two linear equations in two unknowns, x_1 and x_2:

$$2x_1 + 5x_2 = 4$$
$$x_1 + 3x_2 = 5$$

These equations are linear because each additive term in the equation is either a constant (e.g., 4 on the right-hand side of the first equation) or the product of a constant and a variable (e.g., $2x_1$ on the left-hand side of the first equation). Each of the equations $2x_1 + 5x_2 = 4$ and $x_1 + 3x_2 = 5$ literally

represents a line in two-dimensional coordinate space (see Section 2.1.2). Writing the two equations as a matrix equation,

$$\begin{bmatrix} 2 & 5 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 4 \\ 5 \end{bmatrix}$$

$$\underset{(2\times 2)}{\mathbf{A}} \; \underset{(2\times 1)}{\mathbf{x}} = \underset{(2\times 1)}{\mathbf{b}}$$

where

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

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$\mathbf{b} = \begin{bmatrix} 4 \\ 5 \end{bmatrix}$$

The formulation and solution of systems of linear simultaneous equations is taken up in Section 1.4.

1.1.3 Matrix Inverses

In scalar algebra, division is essential to the solution of simple equations. For example,

$$6x = 12$$

$$x = \frac{12}{6} = 2$$

or, equivalently,

$$\frac{1}{6} \times 6x = \frac{1}{6} \times 12$$

$$x = 2$$

where $\frac{1}{6} = 6^{-1}$ is the scalar inverse of 6.

In matrix algebra, there is no direct analog of division, but most square matrices have a *matrix inverse*. The inverse of a square matrix[1] $\mathbf{A}$

[1] It is possible to define various sorts of *generalized inverses* for rectangular matrices and for square matrices that do not have conventional inverses. See Section 1.4.3.

is a square matrix of the same order, written $\mathbf{A}^{-1}$, with the property that $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$. If a square matrix has an inverse, then the matrix is termed *nonsingular*; a square matrix without an inverse is termed *singular*. (When mathematicians first encountered nonzero matrices without inverses, they found the existence of such matrices remarkable or "singular.") If the inverse of a matrix exists, then it is unique; moreover, if for a square matrix $\mathbf{A}$, $\mathbf{AB} = \mathbf{I}$, then necessarily $\mathbf{BA} = \mathbf{I}$, and thus $\mathbf{B} = \mathbf{A}^{-1}$. For example, the inverse of the nonsingular matrix

$$\begin{bmatrix} 2 & 5 \\ 1 & 3 \end{bmatrix}$$

is the matrix

$$\begin{bmatrix} 3 & -5 \\ -1 & 2 \end{bmatrix}$$

as we can readily verify:

$$\begin{bmatrix} 2 & 5 \\ 1 & 3 \end{bmatrix}\begin{bmatrix} 3 & -5 \\ -1 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \checkmark$$

$$\begin{bmatrix} 3 & -5 \\ -1 & 2 \end{bmatrix}\begin{bmatrix} 2 & 5 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \checkmark$$

In scalar algebra, only the number 0 has no inverse. It is simple to show by example that there exist singular *nonzero* matrices: Let us hypothesize that $\mathbf{B}$ is the inverse of the matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

But

$$\mathbf{AB} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} b_{11} & b_{12} \\ 0 & 0 \end{bmatrix} \neq \mathbf{I}_2$$

which contradicts the hypothesis, and $\mathbf{A}$ consequently has no inverse.

There are many methods for finding the inverse of a nonsingular square matrix. I will briefly and informally describe a procedure called *Gaussian elimination* (after the great German mathematician, Carl Friedrich Gauss, 1777–1855). Although there are methods that tend to produce more accurate

numerical results when implemented on a digital computer, elimination has the virtue of relative simplicity, and has applications beyond matrix inversion (as we will see later in this chapter). To illustrate the method of elimination, I will employ the matrix

$$\begin{bmatrix} 2 & -2 & 0 \\ 1 & -1 & 1 \\ 4 & 4 & -4 \end{bmatrix} \tag{1.5}$$

Let us begin by adjoining to this matrix an identity matrix; that is, form the partitioned or *augmented* matrix

$$\left[\begin{array}{ccc|ccc} 2 & -2 & 0 & 1 & 0 & 0 \\ 1 & -1 & 1 & 0 & 1 & 0 \\ 4 & 4 & -4 & 0 & 0 & 1 \end{array}\right]$$

Then attempt to reduce the original matrix to an identity matrix by applying operations of three sorts:

E_I: Multiply each entry in a row of the matrix by a nonzero scalar constant.
E_{II}: Add a scalar multiple of one row to another, replacing the other row.
E_{III}: Exchange two rows of the matrix.

E_I, E_{II}, and E_{III} are called *elementary row operations*.

Starting with the first row, and dealing with each row in turn, insure that there is a nonzero entry in the diagonal position, employing a row interchange for a lower row if necessary. Then divide the row through by its diagonal element (called the *pivot*) to obtain an entry of one in the diagonal position. Finally, add multiples of the current row to the other rows so as to *"sweep out"* the nonzero elements in the pivot column. For the illustration:

1. Divide row 1 by 2,

$$\left[\begin{array}{ccc|ccc} 1 & -1 & 0 & \frac{1}{2} & 0 & 0 \\ 1 & -1 & 1 & 0 & 1 & 0 \\ 4 & 4 & -4 & 0 & 0 & 1 \end{array}\right]$$

2. Subtract the new row 1 from row 2,

$$\left[\begin{array}{ccc|ccc} 1 & -1 & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 1 & -\frac{1}{2} & 1 & 0 \\ 4 & 4 & -4 & 0 & 0 & 1 \end{array}\right]$$

3. Subtract 4 × row 1 from row 3,

$$\left[\begin{array}{ccc|ccc} 1 & -1 & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 1 & -\frac{1}{2} & 1 & 0 \\ 0 & 8 & -4 & -2 & 0 & 1 \end{array}\right]$$

4. Move to row 2; there is a 0 entry in row 2, column 2, so interchange rows 2 and 3,

$$\left[\begin{array}{ccc|ccc} 1 & -1 & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 8 & -4 & -2 & 0 & 1 \\ 0 & 0 & 1 & -\frac{1}{2} & 1 & 0 \end{array}\right]$$

5. Divide row 2 by 8,

$$\left[\begin{array}{ccc|ccc} 1 & -1 & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 1 & -\frac{1}{2} & -\frac{1}{4} & 0 & \frac{1}{8} \\ 0 & 0 & 1 & -\frac{1}{2} & 1 & 0 \end{array}\right]$$

6. Add row 2 to row 1,

$$\left[\begin{array}{ccc|ccc} 1 & 0 & -\frac{1}{2} & \frac{1}{4} & 0 & \frac{1}{8} \\ 0 & 1 & -\frac{1}{2} & -\frac{1}{4} & 0 & \frac{1}{8} \\ 0 & 0 & 1 & -\frac{1}{2} & 1 & 0 \end{array}\right]$$

7. Move to row 3; there is already a 1 in the pivot position; add $\frac{1}{2}\times$ row 3 to row 1,

$$\left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{8} \\ 0 & 1 & -\frac{1}{2} & -\frac{1}{4} & 0 & \frac{1}{8} \\ 0 & 0 & 1 & -\frac{1}{2} & 1 & 0 \end{array}\right]$$

8. Add $\frac{1}{2}\times$ row 3 to row 2,

$$\left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{8} \\ 0 & 1 & 0 & -\frac{1}{2} & \frac{1}{2} & \frac{1}{8} \\ 0 & 0 & 1 & -\frac{1}{2} & 1 & 0 \end{array}\right]$$

Once the original matrix is reduced to the identity matrix, the final columns of the augmented matrix contain the inverse, as we may verify for the example:

$$\begin{bmatrix} 2 & -2 & 0 \\ 1 & -1 & 1 \\ 4 & 4 & -4 \end{bmatrix} \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{8} \\ -\frac{1}{2} & \frac{1}{2} & \frac{1}{8} \\ -\frac{1}{2} & 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \checkmark$$

It is simple to explain why the elimination method works: Each elementary row operation may be represented as multiplication on the left by an appropriately formulated square matrix. Thus, for example, to interchange the second and third rows, we may multiply on the left by

$$\mathbf{E}_{III} \equiv \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

The elimination procedure applies a sequence of (say p) elementary row operations to the augmented matrix $[\underset{(n\times n)}{\mathbf{A}}, \mathbf{I}_n]$, which we may write as

$$\mathbf{E}_p \cdots \mathbf{E}_2\mathbf{E}_1 [\mathbf{A}, \mathbf{I}_n] = [\mathbf{I}_n, \mathbf{B}]$$

using $\mathbf{E}_i$ to represent the ith operation in the sequence. Defining $\mathbf{E} \equiv \mathbf{E}_p \cdots \mathbf{E}_2\mathbf{E}_1$, we have $\mathbf{E}[\mathbf{A}, \mathbf{I}_n] = [\mathbf{I}_n, \mathbf{B}]$; that is, $\mathbf{EA} = \mathbf{I}_n$ (implying that $\mathbf{E} = \mathbf{A}^{-1}$), and $\mathbf{EI}_n = \mathbf{B}$. Consequently, $\mathbf{B} = \mathbf{E} = \mathbf{A}^{-1}$. If $\mathbf{A}$ is singular, then it cannot be reduced to $\mathbf{I}$ by elementary row operations: At some point in the process, we will find that no nonzero pivot is available.

The matrix inverse obeys the following rules:

$$\begin{aligned} \mathbf{I}^{-1} &= \mathbf{I} \\ (\mathbf{A}^{-1})^{-1} &= \mathbf{A} \\ (\mathbf{A}')^{-1} &= (\mathbf{A}^{-1})' \\ (\mathbf{AB})^{-1} &= \mathbf{B}^{-1}\mathbf{A}^{-1} \\ (c\mathbf{A})^{-1} &= c^{-1}\mathbf{A}^{-1} \end{aligned}$$

(where $\mathbf{A}$ and $\mathbf{B}$ are order-n nonsingular matrices, and c is a nonzero scalar). If $\mathbf{D} = \text{diag}(d_1, d_2, \ldots, d_n)$, and if all $d_i \neq 0$, then $\mathbf{D}$ is nonsingular and $\mathbf{D}^{-1} = \text{diag}(1/d_1, 1/d_2, \ldots, 1/d_n)$. Finally, the inverse of a nonsingular symmetric matrix is itself symmetric.

1.1.4 Determinants

Each square matrix $\mathbf{A}$ is associated with a scalar called its *determinant*, written $\det \mathbf{A}$.[2] For a (2×2) matrix $\mathbf{A}$, the determinant is $\det \mathbf{A} = a_{11}a_{22} - a_{12}a_{21}$. For a (3×3) matrix $\mathbf{A}$, the determinant is

$$\begin{aligned}\det \mathbf{A} = {} & a_{11}a_{22}a_{33} - a_{11}a_{23}a_{32} + a_{12}a_{23}a_{31} \\ & - a_{12}a_{21}a_{33} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31}\end{aligned}$$

Although there is a general definition of the determinant of a square matrix of order n, I find it simpler here to define the determinant implicitly by specifying the following properties (or *axioms*):

- **D1:** Multiplying a row of a square matrix by a scalar constant multiplies the determinant of the matrix by the same constant.
- **D2:** Adding a multiple of one row to another leaves the determinant unaltered.
- **D3:** Interchanging two rows changes the sign of the determinant.
- **D4:** $\det \mathbf{I} = 1$.

Axioms D1, D2, and D3 specify the effects on the determinant of the three kinds of elementary row operations. Because the Gaussian elimination method described in Section 1.1.3 reduces a square matrix to the identity matrix, these properties, along with axiom D4, are sufficient for establishing the value of the determinant. Indeed, the determinant is simply the product of the pivot elements, with the sign of the product reversed if, in the course of elimination, an odd number of row interchanges is employed. For the illustrative matrix in Equation 1.5 (on page 13), then, the determinant is $-(2)(8)(1) = -16$, because there was one row interchange (in step 4) and the pivots were 2, 8, and 1 (steps 1, 5, and 7). If a matrix is singular, then one or more of the pivots are zero, and the determinant is zero. Conversely, a nonsingular matrix has a nonzero determinant.

Determinants also occasionally appear directly in statistical applications, for example, in the formula for the multivariate-normal distribution (see Section 3.3.5).

1.1.5 The Kronecker Product

Suppose that $\mathbf{A}$ is an $m \times n$ matrix and that $\mathbf{B}$ is a $p \times q$ matrix. Then the *Kronecker product* of $\mathbf{A}$ and $\mathbf{B}$, denoted $\mathbf{A} \otimes \mathbf{B}$, is defined as

[2] An alternative common notation for $\det \mathbf{A}$ is $|\mathbf{A}|$.

$$\underset{(mp\times nq)}{\mathbf{A}\otimes\mathbf{B}} \equiv \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \cdots & a_{2n}\mathbf{B} \\ \vdots & \vdots & & \vdots \\ a_{m1}\mathbf{B} & a_{m2}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{bmatrix}$$

Named after the 19th-century German mathematician Leopold Kronecker, the Kronecker product is sometimes useful in statistics for compactly representing patterned matrices. For example,

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \otimes \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} = \left[\begin{array}{cc|cc|cc} \sigma_1^2 & \sigma_{12} & 0 & 0 & 0 & 0 \\ \sigma_{12} & \sigma_2^2 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & \sigma_1^2 & \sigma_{12} & 0 & 0 \\ 0 & 0 & \sigma_{12} & \sigma_2^2 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & \sigma_1^2 & \sigma_{12} \\ 0 & 0 & 0 & 0 & \sigma_{12} & \sigma_2^2 \end{array}\right]$$

Many of the properties of the Kronecker product are similar to those of ordinary matrix multiplication; in particular,

$$\mathbf{A}\otimes(\mathbf{B}+\mathbf{C}) = \mathbf{A}\otimes\mathbf{B} + \mathbf{A}\otimes\mathbf{C}$$
$$(\mathbf{B}+\mathbf{C})\otimes\mathbf{A} = \mathbf{B}\otimes\mathbf{A} + \mathbf{C}\otimes\mathbf{A}$$
$$(\mathbf{A}\otimes\mathbf{B})\otimes\mathbf{D} = \mathbf{A}\otimes(\mathbf{B}\otimes\mathbf{D})$$
$$c(\mathbf{A}\otimes\mathbf{B}) = (c\mathbf{A})\otimes\mathbf{B} = \mathbf{A}\otimes(c\mathbf{B})$$

where $\mathbf{B}$ and $\mathbf{C}$ are matrices of the same order, and c is a scalar. As well, like matrix multiplication, the Kronecker product is not commutative: In general, $\mathbf{A}\otimes\mathbf{B} \neq \mathbf{B}\otimes\mathbf{A}$. Additionally, for matrices $\underset{(m\times n)}{\mathbf{A}}$, $\underset{(p\times q)}{\mathbf{B}}$, $\underset{(n\times r)}{\mathbf{C}}$, and $\underset{(q\times s)}{\mathbf{D}}$,

$$(\mathbf{A}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{D}) = \mathbf{AC}\otimes\mathbf{BD}$$

Consequently, if $\underset{(n\times n)}{\mathbf{A}}$ and $\underset{(m\times m)}{\mathbf{B}}$ are nonsingular matrices, then

$$(\mathbf{A}\otimes\mathbf{B})^{-1} = \mathbf{A}^{-1}\otimes\mathbf{B}^{-1}$$

because

$$(\mathbf{A}\otimes\mathbf{B})\left(\mathbf{A}^{-1}\otimes\mathbf{B}^{-1}\right) = (\mathbf{A}\mathbf{A}^{-1})\otimes(\mathbf{B}\mathbf{B}^{-1}) = \mathbf{I}_n\otimes\mathbf{I}_m = \mathbf{I}_{(nm\times nm)}$$

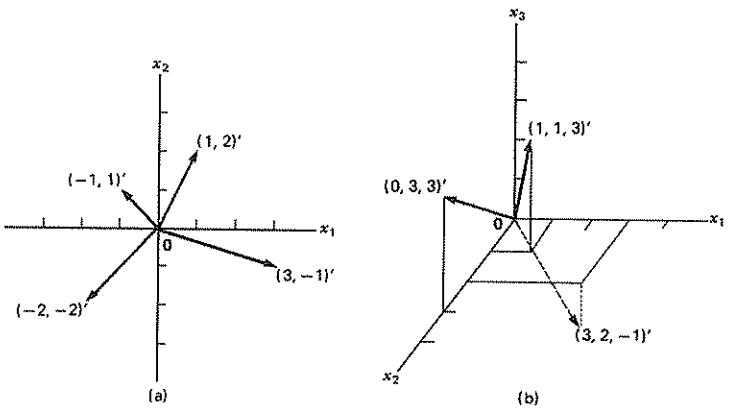

Figure 1.1 Examples of geometric vectors in (a) two-dimensional and (b) three-dimensional space. Each vector is a directed line segment from the origin ($\mathbf{0}$) to the point whose coordinates are given by the entries of the vector.

Finally, for any matrices $\mathbf{A}$ and $\mathbf{B}$,

$$(\mathbf{A} \otimes \mathbf{B})' = \mathbf{A}' \otimes \mathbf{B}'$$

and for square matrices $\mathbf{A}$ and $\mathbf{B}$ of order m and n, respectively,

$$\text{trace}(\mathbf{A} \otimes \mathbf{B}) = \text{trace}(\mathbf{A}) \times \text{trace}(\mathbf{B})$$
$$\det(\mathbf{A} \otimes \mathbf{B}) = (\det \mathbf{A})^m (\det \mathbf{B})^n$$

1.2 Basic Vector Geometry

Considered algebraically, vectors are one-column (or one-row) matrices. Vectors also have the following geometric interpretation: The vector $\mathbf{x} = [x_1, x_2, \ldots, x_n]'$ is represented as a directed line segment extending from the origin of an n-dimensional Cartesian coordinate space to the point defined by the entries (called the *coordinates*) of the vector. Some examples of geometric vectors in two- and three-dimensional space are shown in Figure 1.1.

The basic arithmetic operations defined for vectors have simple geometric interpretations. To add two vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ is, in effect, to place the "tail"

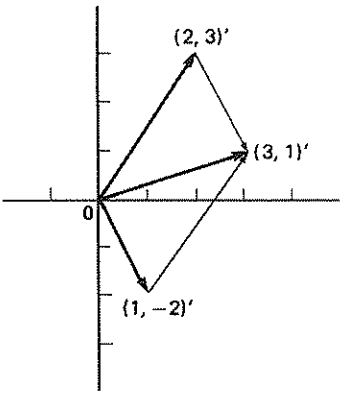

Figure 1.2 Vectors are added by placing the "tail" of one on the tip of the other and completing the parallelogram. The sum is the diagonal of the parallelogram starting at the origin.

of one at the tip of the other. When a vector is shifted from the origin in this manner, it retains its length and orientation (the angles that it makes with respect to the coordinate axes); length and orientation serve to define a vector uniquely. The operation of vector addition, illustrated in two dimensions in Figure 1.2, is equivalent to completing a parallelogram in which $\mathbf{x}_1$ and $\mathbf{x}_2$ are two adjacent sides; the vector sum is the diagonal of the parallelogram, starting at the origin.

As shown in Figure 1.3, the difference $\mathbf{x}_1 - \mathbf{x}_2$ is a vector whose length and orientation are obtained by proceeding from the tip of $\mathbf{x}_2$ to the tip of $\mathbf{x}_1$. Likewise, $\mathbf{x}_2 - \mathbf{x}_1$ proceeds from $\mathbf{x}_1$ to $\mathbf{x}_2$.

The *length* of a vector $\mathbf{x}$, denoted $||\mathbf{x}||$, is the square root of its sum of squared coordinates:

$$\|\mathbf{x}\| = \sqrt{\sum_{i=1}^{n} x_i^2}$$

This result follows from the Pythagorean theorem in two dimensions, as shown in Figure 1.4(a). The result can be extended one dimension at a time to higher-dimensional coordinate spaces, as shown for a three-dimensional space in Figure 1.4(b). The *distance* between two vectors $\mathbf{x}_1$ and $\mathbf{x}_2$, defined as the distance separating their tips, is given by $||\mathbf{x}_1 - \mathbf{x}_2|| = ||\mathbf{x}_2 - \mathbf{x}_1||$ (see Figure 1.3).

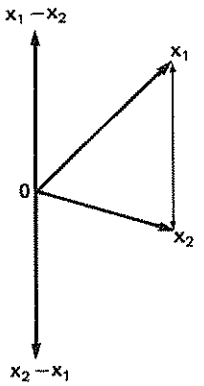

Figure 1.3 Vector differences $\mathbf{x}_1 - \mathbf{x}_2$ and $\mathbf{x}_2 - \mathbf{x}_1$.

The product $a\mathbf{x}$ of a scalar a and a vector $\mathbf{x}$ is a vector of length $|a| \times ||\mathbf{x}||$, as is readily verified:

$$||a\mathbf{x}|| = \sqrt{\sum (ax_i)^2}$$

$$= \sqrt{a^2 \sum x_i^2}$$

$$= |a| \times ||\mathbf{x}||$$

If the scalar a is positive, then the orientation of $a\mathbf{x}$ is the same as that of $\mathbf{x}$; if a is negative, then $a\mathbf{x}$ is *collinear* with (i.e., along the same line as) $\mathbf{x}$ but in the opposite direction. The negative $-\mathbf{x} = (-1)\mathbf{x}$ of $\mathbf{x}$ is, therefore, a vector of the same length as $\mathbf{x}$ but of opposite orientation. These results are illustrated for two dimensions in Figure 1.5.

1.3 Vector Spaces and Subspaces

The *vector space of dimension n* is the infinite set of all vectors $\mathbf{x} = (x_1, x_2, \ldots, x_n)'$; the coordinates x_i may be any real numbers. The vector space of dimension 1 is, therefore, the real line; the vector space of dimension 2 is the plane; and so on.

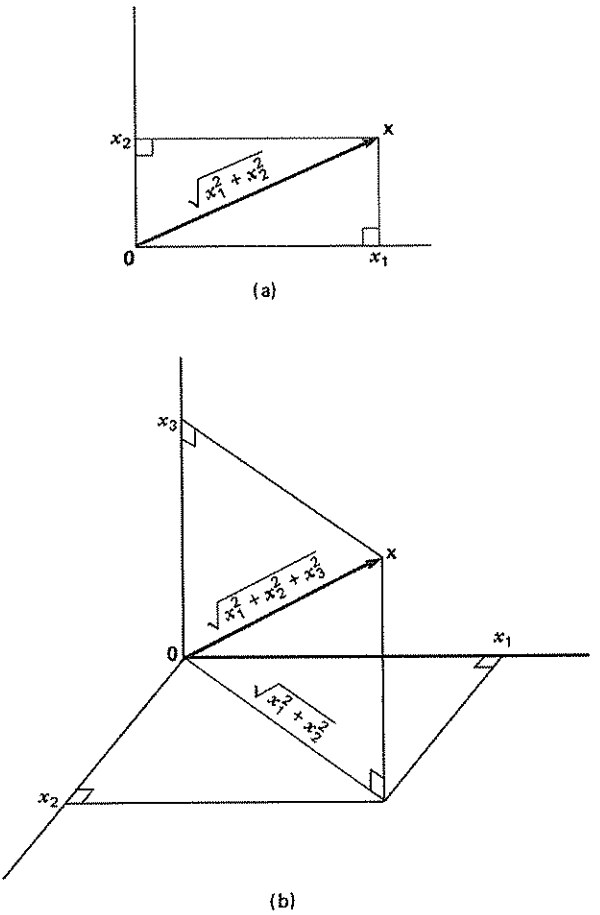

Figure 1.4 The length of a vector is the square root of its sum of squared coordinates, $||\mathbf{x}|| = \sqrt{\sum_{i=1}^{n} x_i^2}$. This result is illustrated in (a) two and (b) three dimensions.

The *subspace* of the n-dimensional vector space that is *generated* by a set of k vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$ is the subset of vectors $\mathbf{y}$ in the space that can be expressed as linear combinations of the generating set:

$$\mathbf{y} = a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \cdots + a_k\mathbf{x}_k$$

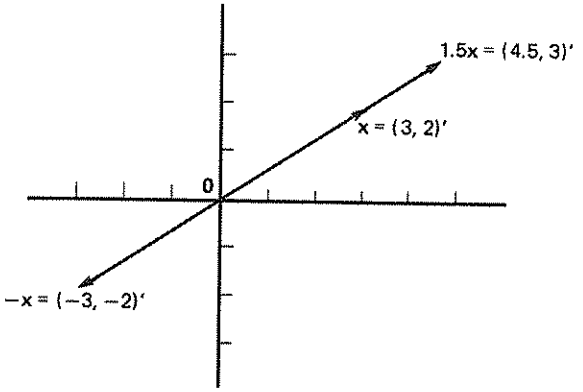

Figure 1.5 Product $a\mathbf{x}$ of a scalar and a vector, illustrated in two dimensions. The vector $a\mathbf{x}$ is collinear with $\mathbf{x}$; it is in the same direction as $\mathbf{x}$ if $a > 0$, and in the opposite direction from $\mathbf{x}$ if $a < 0$.

The set of vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$ is said to *span* the subspace that it generates. Notice that each of $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ is a vector, with n coordinates; that is, $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$ is a set of k vectors, *not* a vector with k coordinates.

A set of vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$ is *linearly independent* if no vector in the set can be expressed as a linear combination of other vectors:

$$\mathbf{x}_j = a_1\mathbf{x}_1 + \cdots + a_{j-1}\mathbf{x}_{j-1} + a_{j+1}\mathbf{x}_{j+1} + \cdots + a_k\mathbf{x}_k \tag{1.6}$$

(where some of the constants a_l can be 0). Equivalently, the set of vectors is linearly independent if there are no constants $b_1, b_2, \ldots, b_k$, not all 0, for which

$$b_1\mathbf{x}_1 + b_2\mathbf{x}_2 + \cdots + b_k\mathbf{x}_k = \underset{(n\times 1)}{\mathbf{0}} \tag{1.7}$$

Equation 1.6 or 1.7 is called a linear *dependency* or *collinearity*. If these equations hold, then the vectors comprise a *linearly dependent* set. Note that the zero vector is linearly dependent on every other vector, inasmuch as $\mathbf{0} = 0\mathbf{x}$.

The *dimension* of the subspace spanned by a set of vectors is the number of vectors in the largest linearly independent subset. The dimension of the subspace spanned by $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$ cannot, therefore, exceed the

smaller of k and n. These relations are illustrated for a vector space of dimension $n = 3$ in Figure 1.6. Figure 1.6(a) shows the one-dimensional subspace (i.e., the line) generated by a single nonzero vector $\mathbf{x}$; Figure 1.6(b) shows the one-dimensional subspace generated by two collinear vectors $\mathbf{x}_1$ and $\mathbf{x}_2$; Figure 1.6(c) shows the two-dimensional subspace (the plane) generated by two linearly independent vectors $\mathbf{x}_1$ and $\mathbf{x}_2$; and Figure 1.6(d) shows the plane generated by three linearly dependent vectors $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$, no two of which are collinear. (In this last case, any one of the three vectors lies in the plane generated by the other two.)

A linearly independent set of vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$—such as $\{\mathbf{x}\}$ in Figure 1.6(a) or $\{\mathbf{x}_1, \mathbf{x}_2\}$ in Figure 1.6(c)—is said to provide a *basis* for the subspace that it spans. Any vector $\mathbf{y}$ in this subspace can be written *uniquely* as a linear combination of the basis vectors:

$$\mathbf{y} = c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \cdots + c_k\mathbf{x}_k$$

The constants $c_1, c_2, \ldots, c_k$ are called the *coordinates of* $\mathbf{y}$ *with respect to the basis* $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$. Because $\mathbf{0} = 0\mathbf{x}_1 + 0\mathbf{x}_2 + \cdots + 0\mathbf{x}_k$, the zero vector is included in every subspace.

The coordinates of a vector with respect to a basis for a two-dimensional subspace can be found geometrically by the parallelogram rule of vector addition, as illustrated in Figure 1.7. Finding coordinates algebraically entails the solution of a system of linear simultaneous equations in which the c_js are the unknowns:

$$\begin{aligned}\underset{(n\times 1)}{\mathbf{y}} &= c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \cdots + c_k\mathbf{x}_k \\ &= [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k]\begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_k \end{bmatrix} \\ &= \underset{(n\times k)}{\mathbf{X}}\ \underset{(k\times 1)}{\mathbf{c}}\end{aligned}$$

When the vectors in $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$ are linearly independent, the matrix $\mathbf{X}$ is of full column rank k, and the equations have a unique solution. The concept of rank and the solution of systems of linear simultaneous equations are taken up in Section 1.4.

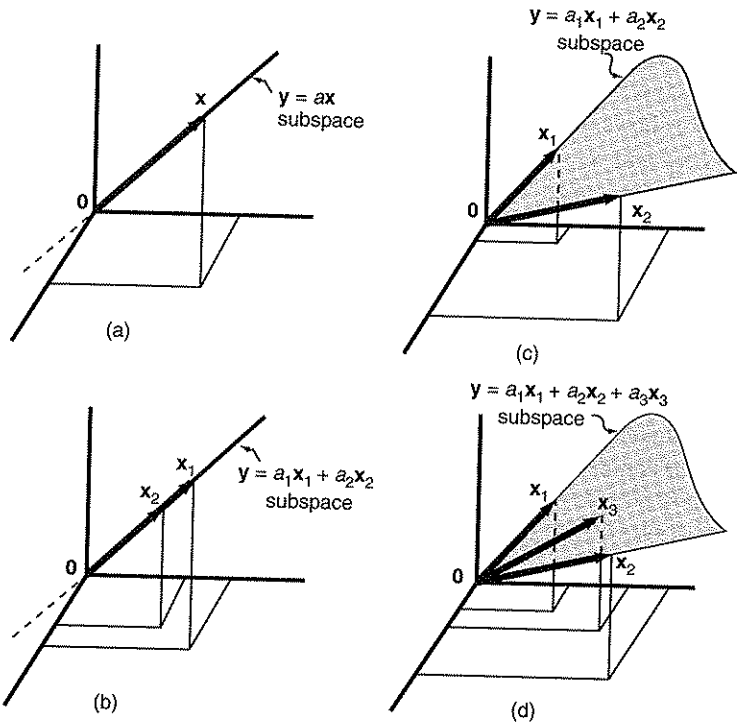

Figure 1.6 Subspaces generated by sets of vectors in three-dimensional space. (a) One nonzero vector generates a one-dimensional subspace (a line). (b) Two collinear vectors also generate a one-dimensional subspace. (c) Two linearly independent vectors generate a two-dimensional subspace (a plane). (d) Three linearly dependent vectors, two of which are linearly independent, generate a two-dimensional subspace. The planes in (c) and (d) extend infinitely; they are drawn between $\mathbf{x}_1$ and $\mathbf{x}_2$ only for clarity.

1.3.1 Orthogonality and Orthogonal Projections

Recall that the inner product of two vectors is the sum of products of their coordinates:

$$\mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^{n} x_i y_i$$

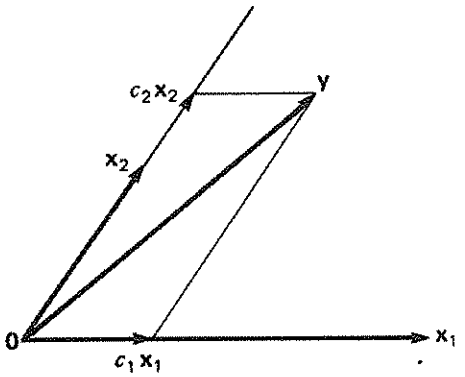

Figure 1.7 The coordinates of $\mathbf{y}$ with respect to the basis $\{\mathbf{x}_1, \mathbf{x}_2\}$ of a two-dimensional subspace can be found from the parallelogram rule of vector addition.

Two vectors $\mathbf{x}$ and $\mathbf{y}$ are *orthogonal* (i.e., perpendicular) if their inner product is 0. The essential geometry of vector orthogonality is shown in Figure 1.8. Although $\mathbf{x}$ and $\mathbf{y}$ lie in an n-dimensional space (and therefore cannot, in general, be visualized directly), they span a subspace of dimension two which, by convention, I make the plane of the paper.[3] When $\mathbf{x}$ and $\mathbf{y}$ are orthogonal [as in Figure 1.8(a)], the two right triangles with vertices $(\mathbf{0}, \mathbf{x}, \mathbf{x} + \mathbf{y})$ and $(\mathbf{0}, \mathbf{x}, \mathbf{x} - \mathbf{y})$ are congruent; consequently, $||\mathbf{x} + \mathbf{y}|| = ||\mathbf{x} - \mathbf{y}||$. Because the squared length of a vector is the inner product of the vector with itself ($\mathbf{x} \cdot \mathbf{x} = \sum x_i^2$), we have

$$
\begin{aligned}
(\mathbf{x} + \mathbf{y}) \cdot (\mathbf{x} + \mathbf{y}) &= (\mathbf{x} - \mathbf{y}) \cdot (\mathbf{x} - \mathbf{y}) \\
\mathbf{x} \cdot \mathbf{x} + 2\mathbf{x} \cdot \mathbf{y} + \mathbf{y} \cdot \mathbf{y} &= \mathbf{x} \cdot \mathbf{x} - 2\mathbf{x} \cdot \mathbf{y} + \mathbf{y} \cdot \mathbf{y} \\
4\mathbf{x} \cdot \mathbf{y} &= 0 \\
\mathbf{x} \cdot \mathbf{y} &= 0
\end{aligned}
$$

When, in contrast, $\mathbf{x}$ and $\mathbf{y}$ are not orthogonal [as in Figure 1.8(b)], then $||\mathbf{x} + \mathbf{y}|| \neq ||\mathbf{x} - \mathbf{y}||$, and $\mathbf{x} \cdot \mathbf{y} \neq 0$.

[3] It is helpful to employ this device in applying vector geometry to statistical problems, where the subspace of interest can often be confined to two or three dimensions, even though the dimension of the full vector space is typically equal to the sample size n.

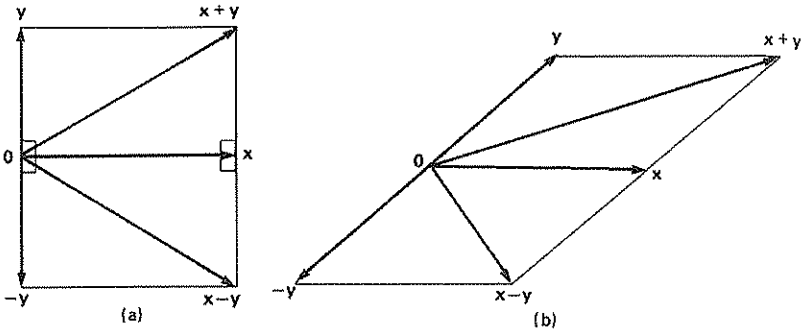

Figure 1.8 When two vectors $\mathbf{x}$ and $\mathbf{y}$ are orthogonal, as in (a), their inner product $\mathbf{x} \cdot \mathbf{y}$ is 0. When the vectors are not orthogonal, as in (b), their inner product is nonzero.

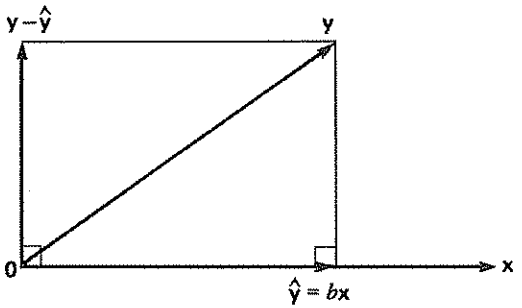

Figure 1.9 The orthogonal projection $\widehat{\mathbf{y}} = b\mathbf{x}$ of $\mathbf{y}$ onto $\mathbf{x}$.

The definition of orthogonality can be extended to matrices in the following manner: The matrix $\underset{(n\times k)}{\mathbf{X}}$ is orthogonal if each pair of its columns is orthogonal—that is, if $\mathbf{X}'\mathbf{X}$ is diagonal.[4] The matrix $\mathbf{X}$ is *orthonormal* if $\mathbf{X}'\mathbf{X} = \mathbf{I}$.

The *orthogonal projection* of one vector $\mathbf{y}$ onto another vector $\mathbf{x}$ is a scalar multiple $\widehat{\mathbf{y}} = b\mathbf{x}$ of $\mathbf{x}$ such that $(\mathbf{y} - \widehat{\mathbf{y}})$ is orthogonal to $\mathbf{x}$. The geometry

[4]The i, jth entry of $\mathbf{X}'\mathbf{X}$ is $\mathbf{x}_i'\mathbf{x}_j = \mathbf{x}_i \cdot \mathbf{x}_j$, where $\mathbf{x}_i$ and $\mathbf{x}_j$ are, respectively, the ith and jth columns of $\mathbf{X}$. The ith diagonal entry of $\mathbf{X}'\mathbf{X}$ is likewise $\mathbf{x}_i'\mathbf{x}_i = \mathbf{x}_i \cdot \mathbf{x}_i$.

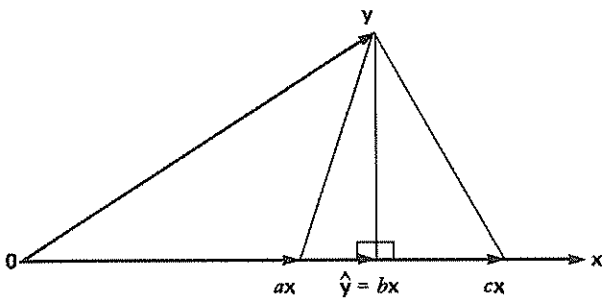

Figure 1.10 The orthogonal projection $\widehat{\mathbf{y}} = b\mathbf{x}$ is the point along the line spanned by $\mathbf{x}$ that is closest to $\mathbf{y}$.

of orthogonal projection is illustrated in Figure 1.9. By the Pythagorean theorem (see Figure 1.10), $\widehat{\mathbf{y}}$ is the point along the line spanned by $\mathbf{x}$ that is closest to $\mathbf{y}$. To find b, we note that

$$\mathbf{x} \cdot (\mathbf{y} - \widehat{\mathbf{y}}) = \mathbf{x} \cdot (\mathbf{y} - b\mathbf{x}) = 0$$

Thus, $\mathbf{x} \cdot \mathbf{y} - b\mathbf{x} \cdot \mathbf{x} = 0$ and $b = (\mathbf{x} \cdot \mathbf{y})/(\mathbf{x} \cdot \mathbf{x})$.

The orthogonal projection of $\mathbf{y}$ onto $\mathbf{x}$ can be used to determine the angle w separating two vectors, by finding its cosine. Because the cosine function is symmetric around $w = 0$, it does not matter in which direction we measure an angle, and I will simply treat angles as positive. The cosine and other basic trigonometric functions are reviewed in Section 2.1.5. I will distinguish between two cases:[5] In Figure 1.11(a), the angle separating the vectors is between 0° and 90°; in Figure 1.11(b), the angle is between 90° and 180°. In the first instance,

$$\cos w = \frac{||\widehat{\mathbf{y}}||}{||\mathbf{y}||} = \frac{b||\mathbf{x}||}{||\mathbf{y}||} = \frac{\mathbf{x} \cdot \mathbf{y}}{||\mathbf{x}||^2} \times \frac{||\mathbf{x}||}{||\mathbf{y}||} = \frac{\mathbf{x} \cdot \mathbf{y}}{||\mathbf{x}|| \times ||\mathbf{y}||}$$

[5] By convention, we examine the smaller of the two angles separating a pair of vectors, and, therefore, never encounter angles that exceed 180°. Call the smaller angle w; then the larger angle is $360 - w$. This convention is of no consequence because $\cos(360 - w) = \cos w$.

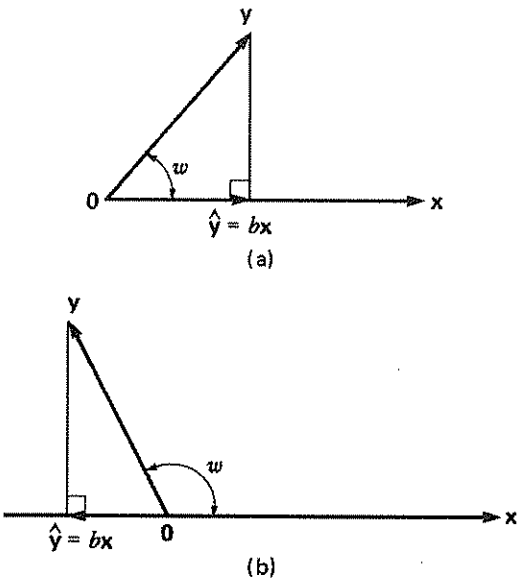

Figure 1.11 The angle w separating two vectors, $\mathbf{x}$ and $\mathbf{y}$: (a) $0° < w < 90°$; (b) $90° < w < 180°$.

and, likewise, in the second instance,

$$\cos w = -\frac{||\widehat{\mathbf{y}}||}{||\mathbf{y}||} = \frac{b||\mathbf{x}||}{||\mathbf{y}||} = \frac{\mathbf{x} \cdot \mathbf{y}}{||\mathbf{x}|| \times ||\mathbf{y}||}$$

In both instances, the sign of b for the orthogonal projection of $\mathbf{y}$ onto $\mathbf{x}$ correctly reflects the sign of $\cos w$.

The orthogonal projection of a vector $\mathbf{y}$ onto the subspace spanned by a set of vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$ is the vector

$$\widehat{\mathbf{y}} = b_1\mathbf{x}_1 + b_2\mathbf{x}_2 + \cdots + b_k\mathbf{x}_k$$

formed as a linear combination of the $\mathbf{x}_j$s such that $(\mathbf{y} - \widehat{\mathbf{y}})$ is orthogonal to each and every vector $\mathbf{x}_j$ in the set. The geometry of orthogonal projection for $k = 2$ is illustrated in Figure 1.12. The vector $\widehat{\mathbf{y}}$ is the point closest to $\mathbf{y}$ in the subspace spanned by the $\mathbf{x}_j$s.

Placing the constants b_j into a vector $\mathbf{b}$, and gathering the vectors $\mathbf{x}_j$ into an $(n \times k)$ matrix $\mathbf{X} \equiv [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k]$, we have $\widehat{\mathbf{y}} = \mathbf{X}\mathbf{b}$. By the definition of an orthogonal projection,

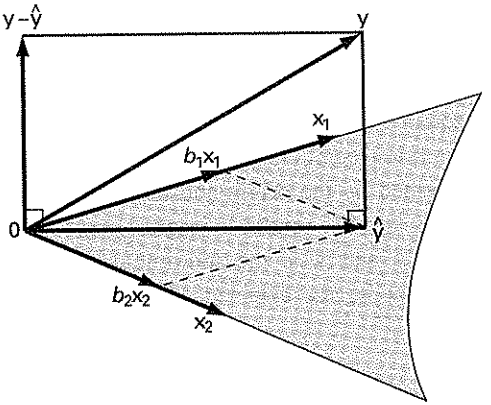

Figure 1.12 The orthogonal projection $\widehat{\mathbf{y}}$ of $\mathbf{y}$ onto the subspace (plane) spanned by $\mathbf{x}_1$ and $\mathbf{x}_2$.

$$\mathbf{x}_j \cdot (\mathbf{y} - \widehat{\mathbf{y}}) = \mathbf{x}_j \cdot (\mathbf{y} - \mathbf{X}\mathbf{b}) = 0 \quad \text{for } j = 1, \ldots, k \tag{1.8}$$

Equivalently, $\mathbf{X}'(\mathbf{y} - \mathbf{X}\mathbf{b}) = \mathbf{0}$, or $\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}\mathbf{b}$. We can solve this matrix equation uniquely for $\mathbf{b}$ as long as $\mathbf{X}'\mathbf{X}$ is nonsingular, in which case $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. The matrix $\mathbf{X}'\mathbf{X}$ is nonsingular if $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$ is a linearly independent set of vectors, providing a basis for the subspace that it generates; otherwise, $\mathbf{b}$ is not unique.

The application of the geometry of orthogonal projections to linear least-squares regression is quite direct. For example, suppose that the vector $\mathbf{x}$ in Figures 1.9 and 1.11 represents the explanatory ("independent") variable in a simple regression; the vector $\mathbf{y}$ represents the response ("dependent") variable; and both variables are expressed as deviations from their means, $\mathbf{x} = \{X_i - \overline{X}\}$, $\mathbf{y} = \{Y_i - \overline{Y}\}$. Then $\widehat{\mathbf{y}} = B\mathbf{x}$ is the mean-deviation vector of fitted ("predicted") Y-values from the linear least-squares regression of Y on X; b is the slope coefficient for the regression; and $\mathbf{y} - \widehat{\mathbf{y}}$ is the vector of least-square residuals. By the Pythagorean theorem,

$$\|\mathbf{y}\|^2 = \|\widehat{\mathbf{y}}\|^2 + \|\mathbf{y} - \widehat{\mathbf{y}}\|^2$$

which shows the decomposition of the total sum of squares for Y into the regression and residual sums of squares—the so-called "analysis of variance" for the regression. The correlation r between X and Y is then the cosine of the angle w separating their mean-deviation vectors.

Suppose similarly that $\mathbf{y}$ is the mean-deviation vector for the response variable and $\mathbf{x}_1$ and $\mathbf{x}_2$ the mean-deviation vectors for two explanatory variables in a multiple regression. Then Figure 1.12 represents the linear least-squares regression of Y on X_1 and X_2; b_1 and b_2 are the partial regression coefficients for the two explanatory variables; the right triangle formed by the origin and the vectors $\mathbf{y}$ and $\widehat{\mathbf{y}}$ gives the analysis of variance for the multiple regression; and the cosine of the angle separating $\mathbf{y}$ and $\widehat{\mathbf{y}}$ is the multiple-correlation coefficient R for the regression—that is, the correlation between observed and fitted Y-values.

1.4 Matrix Rank and the Solution of Linear Simultaneous Equations

1.4.1 Rank

The *row space* of an $(m \times n)$ matrix $\mathbf{A}$ is the subspace of the n-dimensional vector space spanned by the m rows of $\mathbf{A}$ (treated as a set of vectors). The *rank* of $\mathbf{A}$ is the dimension of its row space, that is, the maximum number of linearly independent rows in $\mathbf{A}$. It follows immediately that $\text{rank}(\mathbf{A}) \leq \min(m, n)$.

For example, the row space of the matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

consists of all vectors

$$\begin{aligned} \mathbf{x}' &= a[1, 0, 0] + b[0, 1, 0] \\ &= [a, b, 0] \end{aligned}$$

for any values of a and b. This subspace is of dimension 2, and thus $\text{rank}(\mathbf{A}) = 2$.

A matrix is said to be in *reduced row-echelon form* (*RREF*) if it satisfies the following criteria:

- **R1:** All of its nonzero rows (if any) precede all of its zero rows (if any).
- **R2:** The first nonzero entry (proceeding from left to right) in each nonzero row, called the *leading entry* in the row, is 1.
- **R3:** The leading entry in each nonzero row after the first is to the right of the leading entry in the previous row.
- **R4:** All other entries are 0 in a *column* containing a leading entry.

Reduced row-echelon form is displayed schematically in Equation 1.9, where the asterisks represent elements of arbitrary value:

$$\left[\begin{array}{ccccccccccccccc}
0 & \cdots & 0 & 1 & * & \cdots & * & 0 & * & \cdots & * & 0 & * & \cdots & * \\
0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 1 & * & \cdots & * & 0 & * & \cdots & * \\
\vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & & & \\
0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 1 & * & \cdots & * \\
\hline
0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\
\vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\
0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0
\end{array}\right]$$

nonzero rows

zero rows

(1.9)

The rank of a matrix in RREF is equal to the number of nonzero rows in the matrix: The pattern of leading entries, each located in a column all of whose other elements are zero, insures that no nonzero row can be formed as a linear combination of other rows.

A matrix can be placed in RREF by a sequence of elementary row operations, adapting the elimination procedure first described in Section 1.1.3. For example, starting with the matrix

$$\left[\begin{array}{rrrr} -2 & 0 & -1 & 2 \\ 4 & 0 & 1 & 0 \\ 6 & 0 & 1 & 2 \end{array}\right]$$

1. Divide row 1 by -2,

$$\left[\begin{array}{rrrr} 1 & 0 & \frac{1}{2} & -1 \\ 4 & 0 & 1 & 0 \\ 6 & 0 & 1 & 2 \end{array}\right]$$

2. Subtract $4\times$ row 1 from row 2,

$$\left[\begin{array}{rrrr} 1 & 0 & \frac{1}{2} & -1 \\ 0 & 0 & -1 & 4 \\ 6 & 0 & 1 & 2 \end{array}\right]$$

3. Subtract $6\times$ row 1 from row 3,

$$\left[\begin{array}{rrrr} 1 & 0 & \frac{1}{2} & -1 \\ 0 & 0 & -1 & 4 \\ 0 & 0 & -2 & 8 \end{array}\right]$$

4. Multiply row 2 by -1,

$$\begin{bmatrix} 1 & 0 & \frac{1}{2} & -1 \\ 0 & 0 & 1 & -4 \\ 0 & 0 & -2 & 8 \end{bmatrix}$$

5. Subtract $\frac{1}{2}\times$ row 2 from row 1,

$$\begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & -4 \\ 0 & 0 & -2 & 8 \end{bmatrix}$$

6. Add $2\times$ row 2 to row 3,

$$\begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & -4 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

The rank of a matrix $\mathbf{A}$ is equal to the rank of its RREF $\mathbf{A}_R$, because a zero row in $\mathbf{A}_R$ can only arise if one row of $\mathbf{A}$ is expressible as a linear combination of other rows (or if $\mathbf{A}$ contains a zero row). That is, none of the elementary row operations alters the rank of a matrix. The rank of the matrix transformed to RREF in the example is thus 2.

The RREF of a nonsingular square matrix is the identity matrix, and the rank of a nonsingular square matrix is therefore equal to its order. Conversely, the rank of a singular matrix is less than its order.

I have defined the rank of a matrix $\mathbf{A}$ as the dimension of its row space. It can be shown that the rank of $\mathbf{A}$ is also equal to the dimension of its *column space*—that is, to the maximum number of linearly independent columns in $\mathbf{A}$.

1.4.2 Linear Simultaneous Equations

A system of m linear simultaneous equations in n unknowns can be written in matrix form as

$$\underset{(m\times n)}{\mathbf{A}}\ \underset{(n\times 1)}{\mathbf{x}} = \underset{(m\times 1)}{\mathbf{b}} \tag{1.10}$$

where the elements of the coefficient matrix $\mathbf{A}$ and the right-hand side vector $\mathbf{b}$ are prespecified constants, and $\mathbf{x}$ is the vector of unknowns. Suppose that there is an equal number of equations and unknowns—that is, $m = n$. Then

if the coefficient matrix $\mathbf{A}$ is nonsingular, Equation 1.10 has the *unique solution* $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$.

Alternatively, $\mathbf{A}$ may be singular. Then $\mathbf{A}$ can be transformed to RREF by a sequence of (say, p) elementary row operations, representable as successive multiplication on the left by elementary-row-operation matrices:

$$\mathbf{A}_R = \mathbf{E}_p \cdots \mathbf{E}_2\mathbf{E}_1\mathbf{A} = \mathbf{E}\mathbf{A}$$

Applying these operations to both sides of Equation 1.10 produces

$$\begin{aligned} \mathbf{E}\mathbf{A}\mathbf{x} &= \mathbf{E}\mathbf{b} \\ \mathbf{A}_R\mathbf{x} &= \mathbf{b}_R \end{aligned} \tag{1.11}$$

where $\mathbf{b}_R \equiv \mathbf{E}\mathbf{b}$. Equations 1.10 and 1.11 are *equivalent* in the sense that any solution vector $\mathbf{x} = \mathbf{x}^*$ that satisfies one system also satisfies the other.

Let r represent the rank of $\mathbf{A}$. Because $r < n$ (recall that $\mathbf{A}$ is singular), $\mathbf{A}_R$ contains r nonzero rows and $n - r$ zero rows. If any zero row of $\mathbf{A}_R$ is associated with a nonzero entry (say, b) in $\mathbf{b}_R$, then the system of equations is *inconsistent* or *overdetermined*, for it contains the self-contradictory "equation"

$$0x_1 + 0x_2 + \cdots + 0x_n = b \neq 0$$

If, on the other hand, every zero row of $\mathbf{A}_R$ corresponds to a zero entry in $\mathbf{b}_R$, then the equation system is *consistent*, and there is an infinity of solutions satisfying the system: $n - r$ of the unknowns may be given arbitrary values, which then determine the values of the remaining r unknowns. Under this circumstance, we say that the equation system is *underdetermined*.

Suppose now that there are *fewer* equations than unknowns—that is, $m < n$. Then r is necessarily less than n, and the equations are either overdetermined (if a zero row of $\mathbf{A}_R$ corresponds to a nonzero entry of $\mathbf{b}_R$) or underdetermined (if they are consistent). For example, consider the following system of three equations in four unknowns:

$$\begin{bmatrix} -2 & 0 & -1 & 2 \\ 4 & 0 & 1 & 0 \\ 6 & 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 5 \end{bmatrix}$$

Adjoin the right-hand-side vector to the coefficient matrix,

$$\left[\begin{array}{cccc|c} -2 & 0 & -1 & 2 & 1 \\ 4 & 0 & 1 & 0 & 2 \\ 6 & 0 & 1 & 2 & 5 \end{array}\right]$$

and reduce the coefficient matrix to row-echelon form:

1. Divide row 1 by -2,

$$\left[\begin{array}{cccc|c} 1 & 0 & \frac{1}{2} & -1 & -\frac{1}{2} \\ 4 & 0 & 1 & 0 & 2 \\ 6 & 0 & 1 & 2 & 5 \end{array}\right]$$

2. Subtract 4 $\times$ row 1 from row 2, and subtract 6 $\times$ row 1 from row 3,

$$\left[\begin{array}{cccc|c} 1 & 0 & \frac{1}{2} & -1 & -\frac{1}{2} \\ 0 & 0 & -1 & 4 & 4 \\ 0 & 0 & -2 & 8 & 8 \end{array}\right]$$

3. Multiply row 2 by -1,

$$\left[\begin{array}{cccc|c} 1 & 0 & \frac{1}{2} & -1 & -\frac{1}{2} \\ 0 & 0 & 1 & -4 & -4 \\ 0 & 0 & -2 & 8 & 8 \end{array}\right]$$

4. Subtract $\frac{1}{2}\times$ row 2 from row 1, and add 2 $\times$ row 2 to row 3,

$$\left[\begin{array}{cccc|c} 1\swarrow & 0 & 0 & 1 & \frac{3}{2} \\ 0 & 0 & 1\swarrow & -4 & -4 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

(with the leading entries marked by arrows).

Writing the result as a scalar system of equations, we get

$$\begin{aligned} x_1 + x_4 &= \tfrac{3}{2} \\ x_3 - 4x_4 &= -4 \\ 0x_1 + 0x_2 + 0x_3 + 0x_4 &= 0 \end{aligned}$$

The third equation is uninformative, but it does indicate that the original system of equations is consistent. The first two equations imply that the unknowns x_2 and x_4 can be given arbitrary values (say x_2^* and x_4^*), and the values of x_1 and x_3 (corresponding to the leading entries) follow:

$$x_1 = \tfrac{3}{2} - x_4^*$$
$$x_3 = -4 + 4x_4^*$$

and thus any vector

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} \frac{3}{2} - x_4^* \\ x_2^* \\ -4 + 4x_4^* \\ x_4^* \end{bmatrix}$$

is a solution of the system of equations.

Now consider the system of equations

$$\begin{bmatrix} -2 & 0 & -1 & 2 \\ 4 & 0 & 1 & 0 \\ 6 & 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}$$

Attaching $\mathbf{b}$ to $\mathbf{A}$ and transforming the coefficient matrix to RREF yields

$$\left[\begin{array}{cccc|c} 1 & 0 & 0 & 1 & \frac{1}{2} \\ 0 & 0 & 1 & -4 & -2 \\ 0 & 0 & 0 & 0 & 2 \end{array}\right]$$

The last equation,

$$0x_1 + 0x_2 + 0x_3 + 0x_4 = 2$$

is contradictory, implying that the original system of equations has no solution.

Suppose, finally, that there are *more* equations than unknowns: $m > r$. If $\mathbf{A}$ is of full-column rank (i.e., if $r = n$), then $\mathbf{A}_R$ consists of the order-n identity matrix followed by $m - r$ zero rows. If the equations are consistent, they therefore have a unique solution; otherwise, of course, they

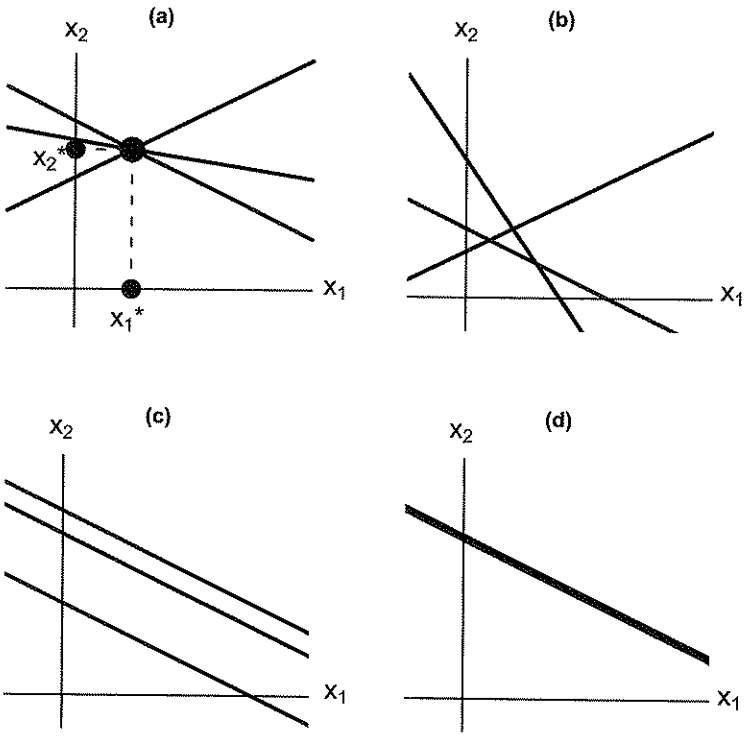

Figure 1.13 Three linear equations in two unknowns: (a) unique solution; (b) and (c) overdetermined; (d) underdetermined (three coincident lines).

are overdetermined. If $r < n$, the equations are either overdetermined (if inconsistent) or underdetermined (if consistent).

To illustrate these results geometrically, consider a system of three linear equations in two unknowns:[6]

$$a_{11}x_1 + a_{12}x_2 = b_1$$
$$a_{21}x_1 + a_{22}x_2 = b_2$$
$$a_{31}x_1 + a_{32}x_2 = b_3$$

[6]The geometric representation of linear equations by lines (or, more generally, by linear surfaces) should. not be confused with the geometric vector representation discussed in Section 1.2.

Each equation describes a line in a two-dimensional coordinate space in which the unknowns define the axes, as illustrated schematically in Figure 1.13. (See Section 2.1.2 for a review of the graphs of linear equations in two and three dimensions.) If the three lines intersect at a point, as in Figure 1.13(a), then there is a *unique solution* to the equation system: Only the pair of values (x_1^*, x_2^*) simultaneously satisfies all three equations. If the three lines fail to intersect at a common point, as in Figures 1.13(b) and (c), then *no* pair of values of the unknowns simultaneously satisfies the three equations, which therefore are overdetermined. Lastly, if the three lines are coincident, as in Figure 1.13(d), then *any* pair of values on the common line satisfies all three equations, and the equations are underdetermined.

When the right-hand side vector $\mathbf{b}$ in a system of linear simultaneous equations is the zero vector, the system of equations is said to be *homogeneous*:

$$\underset{(m\times n)}{\mathbf{A}}\ \underset{(n\times 1)}{\mathbf{x}} = \underset{(m\times 1)}{\mathbf{0}}$$

The *trivial solution* $\mathbf{x} = \mathbf{0}$ always satisfies a homogeneous system which, consequently, cannot be inconsistent. From the previous work in this section, we can see that nontrivial solutions exist if $\text{rank}(\mathbf{A}) < n$—that is, when the system is underdetermined.

The results concerning the solution of linear simultaneous equations developed in this section are summarized in Table 1.1.[7]

Linear simultaneous equations have many statistical applications, such as solving for least-squares coefficients in regression analysis.

1.4.3 Generalized Inverses

As explained in Section 1.1.3, only square nonsingular matrices have inverses. All matrices, however—including singular and rectangular matrices—have *generalized inverses*, which are occasionally employed in statistical applications, such as some presentations of linear statistical models.[8]

[7] A note on terminology: Some authors call *any* equation system with more equations than unknowns—whether consistent or not, and regardless of the rank of the coefficient matrix—"overdetermined," and any system with more unknowns than equations "underdetermined." I believe that my usage (which follows, e.g., Davis, 1973) is more natural.

[8] For an extensive discussion of the role of generalized inverses in statistics, see Rao and Mitra (1971).

TABLE 1.1
Solutions of m Linear Simultaneous Equations in n Unknowns

Number of Equations	$m < n$	$m = n$		$m > n$	
Rank of Coefficient Matrix	$r < n$	$r < n$	$r = n$	$r < n$	$r = n$
			General Equation System		
Consistent	Under-determined	Under-determined	Unique solution	Under-determined	Unique solution
Inconsistent	Over-determined	Over-determined	—	Over-determined	Over-determined
			Homogeneous Equation System		
Consistent	Nontrivial solutions	Nontrivial solutions	Trivial solution	Nontrivial solutions	Trivial solution

A generalized inverse of the $(m \times n)$ matrix $\mathbf{A}$ is an $(n \times m)$ matrix $\mathbf{A}^-$ that satisfies the equation

$$\mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A} \tag{1.12}$$

Notice that $\mathbf{A}^-$ is *a* generalized inverse, not *the* generalized inverse of $\mathbf{A}$: Unless $\mathbf{A}$ is square and nonsingular (in which case $\mathbf{A}^- = \mathbf{A}^{-1}$), the generalized inverse is not unique.[9]

There are many ways to find a generalized inverse of a matrix. We can, for example, proceed by Gaussian elimination. Suppose that we begin by putting the matrix $\mathbf{A}$ in RREF by a sequence of elementary row operations; we know that we can represent this process by successive multiplication on the left by suitably configured elementary-row-operations matrices (see Section 1.1.3):

$$\mathbf{E}\mathbf{A} = \mathbf{E}_p \cdots \mathbf{E}_2\mathbf{E}_1\mathbf{A} = \mathbf{A}_R \tag{1.13}$$

where $\mathbf{E} \equiv \mathbf{E}_p \cdots \mathbf{E}_2\mathbf{E}_1$ is a nonsingular $(m \times m)$ matrix. Applying an analogous series of type II and III *elementary column operations* (pivoting

[9]The generalized inverse can be made unique by placing additional restrictions on it beyond Equation 1.12: For example, the *Moore-Penrose generalized inverse* $\mathbf{A}^+$ satisfies four conditions: $\mathbf{A}\mathbf{A}^+\mathbf{A} = \mathbf{A}$; $\mathbf{A}^+\mathbf{A}\mathbf{A}^+ = \mathbf{A}^+$; $\mathbf{A}\mathbf{A}^+$ is symmetric; and $\mathbf{A}^+\mathbf{A}$ is symmetric. In a typical statistical application, however, one generalized inverse is as good as another.

is unnecessary because all of the leading entries in $\mathbf{A}_R$ are already 1), we can further reduce A_R to the following *canonical form*:

$$\underset{(m\times n)}{\mathbf{A}_C} \equiv \mathbf{A}_R\mathbf{E}^* = \mathbf{A}_R\mathbf{E}_1^*\mathbf{E}_2^*\cdots\mathbf{E}_q^* = \begin{bmatrix} \mathbf{I}_r & \underset{(r\times n-r)}{\mathbf{0}} \\ \underset{(m-r\times r)}{\mathbf{0}} & \underset{(m-r\times n-r)}{\mathbf{0}} \end{bmatrix} \tag{1.14}$$

where $\mathbf{E}^* \equiv \mathbf{E}_1^*\mathbf{E}_2^*\cdots\mathbf{E}_q^*$ is a nonsingular $(n \times n)$ matrix; the order r of the identity matrix in the upper-left corner is the rank of $\mathbf{A}$; and any and all of the zero matrices may be absent. For example, if $\mathbf{A}$ is a nonsingular matrix of order n then $r = n$ and none of the zero matrices are required.

Putting together Equations 1.13 and 1.14, we have

$$\mathbf{A}_C = \mathbf{E}\mathbf{A}\mathbf{E}^* \tag{1.15}$$

A generalized inverse of $\mathbf{A}$ is then given by[10]

$$\mathbf{A}^- \equiv \mathbf{E}^*\mathbf{A}_C'\mathbf{E}$$

Consider, for example, the matrix

$$\mathbf{A} = \begin{bmatrix} -2 & 0 & -1 & 2 \\ 4 & 0 & 1 & 0 \\ 6 & 0 & 1 & 2 \end{bmatrix}$$

In Section 1.4.1, I reduced this matrix to row-echelon form by a sequence of elementary row operations:

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

[10]The following proof is adapted from Healy (1986, p. 40): First, $\mathbf{A}_C'$ is a generalized inverse of $\mathbf{A}_C$ (check it!); second, solving Equation 1.15 for $\mathbf{A}$ produces $\mathbf{A} = \mathbf{E}^{-1}\mathbf{A}_C\mathbf{E}^{*-1}$. Then,

$$\begin{aligned} \mathbf{A}\mathbf{A}^-\mathbf{A} &= (\mathbf{E}^{-1}\mathbf{A}_C\mathbf{E}^{*-1})(\mathbf{E}^*\mathbf{A}_C'\mathbf{E})(\mathbf{E}^{-1}\mathbf{A}_C\mathbf{E}^{*-1}) \\ &= \mathbf{E}^{-1}\mathbf{A}_C\mathbf{A}_C'\mathbf{A}_C\mathbf{E}^{*-1} \\ &= \mathbf{E}^{-1}\mathbf{A}_C\mathbf{E}^{*-1} \\ &= \mathbf{A} \end{aligned}$$

which establishes the result.

The reduction to canonical form is completed by exchanging columns 2 and 3, and then sweeping out the fourth column, producing

$$\mathbf{A}_C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Collecting the elementary row and column operations into matrices, we have

$$\mathbf{E} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ -2 & -1 & 0 \\ -1 & -2 & 1 \end{bmatrix}$$

$$\mathbf{E}^* = \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 4 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

from which

$$\begin{aligned} \mathbf{A}^- &= \mathbf{E}^* \mathbf{A}_C' \mathbf{E} \\ &= \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 4 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ -2 & -1 & 0 \\ -1 & -2 & 1 \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 \\ -2 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \end{aligned}$$

is a generalized inverse of $\mathbf{A}$ (as the reader can verify).

Consider a system of m linear simultaneous equations in n unknowns,

$$\underset{(m\times n)}{\mathbf{A}} \; \underset{(n\times 1)}{\mathbf{x}} = \underset{(m\times 1)}{\mathbf{b}}$$

as discussed in Section 1.4.2, and suppose that the system of equations is consistent and underdetermined. Then

$$\mathbf{x}^* = \mathbf{A}^- \mathbf{b} \tag{1.16}$$

provides an arbitrary solution to the equations. If the equation system has a unique solution, then Equation 1.16 yields it. Finally, if the equation

system is overdetermined, then the "solution" provided by Equation 1.16 will fail to satisfy the original system of equations. Thus, if the equation system is consistent, then $\mathbf{A}\mathbf{A}^{-}\mathbf{b} = \mathbf{b}$, and if the system is inconsistent, then $\mathbf{A}\mathbf{A}^{-}\mathbf{b} \neq \mathbf{b}$. The reader may wish to apply these results to the examples in Section 1.4.2.

1.5 Eigenvalues and Eigenvectors

If $\mathbf{A}$ is an order-n square matrix, then the homogeneous system of linear equations

$$(\mathbf{A} - \lambda\mathbf{I}_n)\mathbf{x} = \mathbf{0} \tag{1.17}$$

will have nontrivial solutions only for certain values of the scalar λ. The results in the preceding section suggest that nontrivial solutions exist when the matrix $(\mathbf{A} - \lambda\mathbf{I}_n)$ is singular, that is, when

$$\det(\mathbf{A} - \lambda\mathbf{I}_n) = 0 \tag{1.18}$$

Equation 1.18 is called the *characteristic equation* of the matrix $\mathbf{A}$, and the values of λ for which this equation holds are called the *eigenvalues*, *characteristic roots*, or *latent roots* of $\mathbf{A}$. A vector $\mathbf{x}_1$ satisfying Equation 1.17 for a particular eigenvalue λ_1 is called an *eigenvector*, *characteristic vector*, or *latent vector* of $\mathbf{A}$ associated with λ_1.

Because of its simplicity and straightforward extension, I will examine the (2×2) case in some detail. For this case, the characteristic equation is

$$\det\begin{bmatrix} a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{bmatrix} = 0$$

$$(a_{11} - \lambda)(a_{22} - \lambda) - a_{12}a_{21} = 0$$

$$\lambda^2 - (a_{11} + a_{22})\lambda + a_{11}a_{22} - a_{12}a_{21} = 0$$

Using the quadratic formula to solve the characteristic equation produces the two roots[11]

[11] Review of the *quadratic formula*: The values of x that satisfy the quadratic equation

$$ax^2 + bx + c = 0$$

where a, b, and c are specific constants, are

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

$$\lambda_1 = \frac{1}{2}\left[a_{11} + a_{22} + \sqrt{(a_{11} + a_{22})^2 - 4(a_{11}a_{22} - a_{12}a_{21})}\right] \quad (1.19)$$

$$\lambda_2 = \frac{1}{2}\left[a_{11} + a_{22} - \sqrt{(a_{11} + a_{22})^2 - 4(a_{11}a_{22} - a_{12}a_{21})}\right]$$

These roots are real if the quantity under the radical is non-negative. Notice, incidentally, that $\lambda_1 + \lambda_2 = a_{11} + a_{22}$ (the sum of the eigenvalues of $\mathbf{A}$ is the trace of $\mathbf{A}$), and that $\lambda_1\lambda_2 = a_{11}a_{22} - a_{12}a_{21}$ (the product of the eigenvalues is the determinant of $\mathbf{A}$). Furthermore, if $\mathbf{A}$ is singular, then λ_2 is 0.

If $\mathbf{A}$ is symmetric (as is the case for most statistical applications of eigenvalues and eigenvectors), then $a_{12} = a_{21}$, and Equation 1.19 becomes

$$\lambda_1 = \frac{1}{2}\left[a_{11} + a_{22} + \sqrt{(a_{11} - a_{22})^2 + 4a_{12}^2}\right] \quad (1.20)$$

$$\lambda_2 = \frac{1}{2}\left[a_{11} + a_{22} - \sqrt{(a_{11} - a_{22})^2 + 4a_{12}^2}\right]$$

The eigenvalues of a (2×2) symmetric matrix are necessarily real because the quantity under the radical in Equation 1.20 is the sum of two squares, which cannot be negative.

I will use the following (2×2) matrix as an illustration:

$$\begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

Here,

$$\lambda_1 = \frac{1}{2}\left[1 + 1 + \sqrt{(1-1)^2 + 4(0.5)^2}\right] = 1.5$$

$$\lambda_2 = \frac{1}{2}\left[1 + 1 - \sqrt{(1-1)^2 + 4(0.5)^2}\right] = 0.5$$

To find the eigenvectors associated with $\lambda_1 = 1.5$, solve the homogeneous system of equations

$$\begin{bmatrix} 1 - 1.5 & 0.5 \\ 0.5 & 1 - 1.5 \end{bmatrix}\begin{bmatrix} x_{11} \\ x_{21} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} -0.5 & 0.5 \\ 0.5 & -0.5 \end{bmatrix}\begin{bmatrix} x_{11} \\ x_{21} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

yielding

$$\mathbf{x}_1 = \begin{bmatrix} x_{11} \\ x_{21} \end{bmatrix} = \begin{bmatrix} x_{21}^* \\ x_{21}^* \end{bmatrix}$$

(i.e., any vector with two equal entries). Similarly, for $\lambda_2 = 0.5$, solve

$$\begin{bmatrix} 1-0.5 & 0.5 \\ 0.5 & 1-0.5 \end{bmatrix} \begin{bmatrix} x_{12} \\ x_{22} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix} \begin{bmatrix} x_{12} \\ x_{22} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

which produces

$$\mathbf{x}_2 = \begin{bmatrix} x_{12} \\ x_{22} \end{bmatrix} = \begin{bmatrix} -x_{22}^* \\ x_{22}^* \end{bmatrix}$$

(i.e., any vector whose two entries are the negative of each other). The set of eigenvalues associated with each eigenvector therefore spans a one-dimensional subspace: When one of the entries of the eigenvector is specified, the other entry follows. Notice further that the eigenvectors $\mathbf{x}_1$ and $\mathbf{x}_2$ are orthogonal:

$$\mathbf{x}_1 \cdot \mathbf{x}_2 = -x_{21}^* x_{22}^* + x_{21}^* x_{22}^* = 0$$

Many of the properties of eigenvalues and eigenvectors of (2×2) matrices generalize to $(n \times n)$ matrices. In particular:

- The characteristic equation, $\det(\mathbf{A} - \lambda \mathbf{I}_n) = 0$, of an $(n \times n)$ matrix is an nth-order polynomial in λ; there are, consequently, n eigenvalues, not all necessarily distinct.[12]
- The sum of the eigenvalues of $\mathbf{A}$ is the trace of $\mathbf{A}$.
- The product of the eigenvalues of $\mathbf{A}$ is the determinant of $\mathbf{A}$.
- The number of nonzero eigenvalues of $\mathbf{A}$ is the rank of $\mathbf{A}$.
- A singular matrix, therefore, has a least one zero eigenvalue.
- If $\mathbf{A}$ is a symmetric matrix, then the eigenvalues of $\mathbf{A}$ are all real numbers.
- If the eigenvalues of $\mathbf{A}$ are distinct (i.e., all different), then the set of eigenvectors associated with a particular eigenvalue spans a one-dimensional subspace. If, alternatively, k eigenvalues are equal, then their common set of eigenvectors spans a subspace of dimension k.
- Eigenvectors associated with different eigenvalues are orthogonal.

[12] Finding eigenvalues by solving the characteristic equation directly is not generally an attractive approach, and other, more practical, methods exist for finding eigenvalues and their associated eigenvectors.

Suppose that $\mathbf{A}$ is an $(n \times n)$ symmetric matrix of rank r. Let $\mathbf{\Lambda} \equiv \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_r)$ collect the nonzero eigenvalues of $\mathbf{A}$; let $\mathbf{x}_j$ represent an eigenvector corresponding to λ_j, normed to unit length, $\|\mathbf{x}_j\| = 1$; and let $\mathbf{X} \equiv [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_r]$ collect these eigenvectors. Then

$$\begin{aligned}\mathbf{A} &= \lambda_1 \mathbf{x}_1 \mathbf{x}_1' + \lambda_2 \mathbf{x}_2 \mathbf{x}_2' + \cdots + \lambda_r \mathbf{x}_r \mathbf{x}_r' \\ &= \mathbf{X}\mathbf{\Lambda}\mathbf{X}'\end{aligned} \tag{1.21}$$

Equation 1.21, called the *spectral decomposition* of the matrix $\mathbf{A}$, is the mathematical basis of such statistical techniques as principal components analysis and factor analysis.

Eigenvectors and eigenvalues can be generalized in the following manner: Suppose that $\mathbf{A}$ is an $(n \times n)$ symmetric matrix, and let us replace Equation 1.17 (page 41) with

$$(\mathbf{A} - \lambda \mathbf{B})\mathbf{x} = \mathbf{0}$$

where $\mathbf{B}$ is an $(n \times n)$ symmetric, positive-definite matrix. (See the folllowing section for the definition of positive-definiteness.) Then the values of λ that satisfy this equation are called *generalized eigenvalues of* $\mathbf{A}$ *in the metric of* $\mathbf{B}$, and the corresponding vectors $\mathbf{x}$ are *generalized eigenvectors*. It turns out that the generalized eigenvalues are the ordinary eigenvalues of $\mathbf{A}\mathbf{B}^{-1}$. Generalized eigenvalues and eigenvectors are useful in certain areas of multivariate statistics, such as for hypothesis tests in the multivariate linear model.

Still another generalization of eigenvalues and eigenvectors is to rectangular matrices. Suppose, now, that $\mathbf{A}$ is an $(m \times n)$ matrix of rank r. Then $\mathbf{A}$ can be factored as

$$\underset{(m\times n)}{\mathbf{A}} = \underset{(m\times m)}{\mathbf{B}} \begin{bmatrix} \underset{(r\times r)}{\mathbf{\Lambda}} & \underset{(r\times n-r)}{\mathbf{0}} \\ \underset{(m-r\times r)}{\mathbf{0}} & \underset{(m-r\times n-r))}{\mathbf{0}} \end{bmatrix} \underset{(n\times n)}{\mathbf{C}'} \tag{1.22}$$

where

- $\mathbf{B}$ and $\mathbf{C}$ are orthogonal matrices, and are not generally unique (orthogonal matrices are defined in Section 1.3.1);
- $\mathbf{\Lambda}^2$ is a diagonal matrix containing the nonzero eigenvalues of the matrices $\mathbf{A}'\mathbf{A}$ and $\mathbf{A}\mathbf{A}'$ (which share the same eigenvalues); and
- not all of the zero matrices may be needed. (Indeed, if $r = m = n$, then Equation 1.22 reduces to the spectral decomposition in Equation 1.21.)

Equation 1.22 is termed the *singular-value decomposition* of the matrix $\mathbf{A}$, and the diagonal entries of $\mathbf{\Lambda}$ are the *singular values* of $\mathbf{A}$ (which are, therefore, the square roots of the eigenvalues of $\mathbf{A}'\mathbf{A}$ and $\mathbf{A}\mathbf{A}'$). The singular-value decomposition is useful, for example, for improving the computational efficiency and precision of least-squares calculations.

1.6 Quadratic Forms and Positive-Definite Matrices

The expression

$$\underset{(1\times n)}{\mathbf{x}'}\ \underset{(n\times n)}{\mathbf{A}}\ \underset{(n\times 1)}{\mathbf{x}} \tag{1.23}$$

is called a *quadratic form* in $\mathbf{x}$. In this section (as in typical statistical applications), $\mathbf{A}$ will always be a symmetric matrix. $\mathbf{A}$ is said to be *positive-definite* if the quadratic form in Equation 1.23 is positive for all nonzero $\mathbf{x}$. $\mathbf{A}$ is *positive-semidefinite* if the quadratic form is non-negative (i.e., positive or zero) for all nonzero vectors $\mathbf{x}$. The eigenvalues of a positive-definite matrix are all positive (and, consequently, the matrix is nonsingular); those of a positive-semidefinite matrix are all positive or zero.

Let

$$\underset{(m\times m)}{\mathbf{C}} = \underset{(m\times n)}{\mathbf{B}'}\ \underset{(n\times n)}{\mathbf{A}}\ \underset{(n\times m)}{\mathbf{B}}$$

where $\mathbf{A}$ is positive-definite and $\mathbf{B}$ is of full-column rank $m \leq n$. I will show that $\mathbf{C}$ is also positive-definite. Note, first, that $\mathbf{C}$ is symmetric:

$$\mathbf{C}' = (\mathbf{B}'\mathbf{A}\mathbf{B})' = \mathbf{B}'\mathbf{A}'\mathbf{B} = \mathbf{B}'\mathbf{A}\mathbf{B} = \mathbf{C}$$

If $\mathbf{y}$ is any $(m \times 1)$ nonzero vector, then $\underset{(n\times 1)}{\mathbf{x}} = \mathbf{B}\mathbf{y}$ is also nonzero: Because $\mathbf{B}$ is of rank m, we can select m linearly independent rows from $\mathbf{B}$, forming the nonsingular matrix $\mathbf{B}^*$. Then $\underset{(m\times 1)}{\mathbf{x}^*} = \mathbf{B}^*\mathbf{y}$, which contains a subset of the entries in $\mathbf{x}$, is nonzero because $\mathbf{y} = \mathbf{B}^{*-1}\mathbf{x}^* \neq \mathbf{0}$. Consequently

$$\mathbf{y}'\mathbf{C}\mathbf{y} = \mathbf{y}'\mathbf{B}'\mathbf{A}\mathbf{B}\mathbf{y} = \mathbf{x}'\mathbf{A}\mathbf{x}$$

is necessarily positive, and $\mathbf{C}$ is positive-definite. By similar reasoning, if $\text{rank}(\mathbf{B}) < m$, then $\mathbf{C}$ is positive-semidefinite. The matrix

$\underset{(m\times n)}{\mathbf{B}'}\underset{(n\times m)}{\mathbf{B}} = \mathbf{B}'\mathbf{I}_n\mathbf{B}$ is therefore positive-definite if $\mathbf{B}$ is of full-column rank (because $\mathbf{I}_n$ is clearly positive-definite), and positive-semidefinite otherwise. (Cf., the geometric discussion following Equation 1.8 on page 29.)

Positive-definite and -semidefinite matrices—such as variance-covariance matrices, correlation matrices, and matrices of sums of squares and products—play a prominent role in statistics.

1.6.1 The Cholesky Decomposition

Every $(n \times n)$ symmetric positive-definite matrix $\mathbf{A}$ can be factored uniquely as $\mathbf{A} = \mathbf{U}'\mathbf{U}$, where $\mathbf{U}$ is an upper-triangular matrix with positive diagonal elements; $\mathbf{U}$, called the *Cholesky factor* of $\mathbf{A}$, may be thought of as a kind of matrix square root (though not in the sense developed in Section 1.1.2). The Cholesky factor is named after the 19th-century French mathematician André-Louis Cholesky.

Consider, for example, the (3×3) matrix

$$\mathbf{A} = \begin{bmatrix} 1.0 & 0.5 & 0.3 \\ 0.5 & 1.0 & 0.5 \\ 0.3 & 0.5 & 1.0 \end{bmatrix}$$

and let

$$\mathbf{U} = \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix}$$

represent the Cholesky factor of $\mathbf{A}$. Then,

$$\mathbf{U}'\mathbf{U} = \begin{bmatrix} u_{11}^2 & u_{11}u_{12} & u_{11}u_{13} \\ u_{12}u_{11} & u_{12}^2 + u_{22}^2 & u_{12}u_{13} + u_{22}u_{23} \\ u_{13}u_{11} & u_{13}u_{12} + u_{23}u_{22} & u_{13}^2 + u_{23}^2 + u_{33}^2 \end{bmatrix}$$

$$= \begin{bmatrix} 1.0 & 0.5 & 0.3 \\ 0.5 & 1.0 & 0.5 \\ 0.3 & 0.5 & 1.0 \end{bmatrix} = \mathbf{A}$$

from which

$$u_{11}^2 = 1.0 \Longrightarrow u_{11} = 1.0$$
$$u_{12}u_{11} = u_{12} \times 1 = 0.5 \Longrightarrow u_{12} = 0.5$$

$$u_{12}^2 + u_{22}^2 = 0.5^2 + u_{22}^2 = 1 \Longrightarrow u_{22} = \sqrt{1 - 0.5^2}$$
$$= 0.8660$$
$$u_{13}u_{11} = u_{13} \times 1 = 0.3 \Longrightarrow u_{13} = 0.3$$
$$u_{13}u_{12} + u_{23}u_{22} = 0.3 \times 0.5 + u_{23} \times 0.8660 = 0.5 \Longrightarrow$$
$$u_{23} = (0.5 - 0.3 \times 0.5)/0.8660 = 0.4041$$
$$u_{13}^2 + u_{23}^2 + u_{33}^2 = 0.3^2 + 0.4141^2 + u_{33}^2 = 1 \Longrightarrow$$
$$u_{33} = \sqrt{1 - 0.3^2 - 0.4041^2} = 0.8641$$

and thus

$$\mathbf{U} = \begin{bmatrix} 1.0 & 0.5 & 0.3 \\ 0 & 0.8660 & 0.4041 \\ 0 & 0 & 0.8641 \end{bmatrix}$$

This procedure can be extended to symmetric positive-definite matrices of any order.[13]

1.7 Recommended Reading

There is a plethora of books on linear algebra and matrices. Most presentations develop the fundamental properties of vector spaces, but often, unfortunately, without explicit visual representation.

- Several matrix texts, including Healy (1986), Graybill (1983), Searle (1982), and Green and Carroll (1976), focus specifically on statistical applications. The last of these sources has a strongly geometric orientation.
- Davis (1973), who presents a particularly lucid and simple treatment of matrix algebra, includes some material on vector geometry (limited, however, to two dimensions).
- Namboodiri (1984) provides a compact introduction to matrix algebra (but not to vector geometry).
- Texts on statistical computing, such as Kennedy and Gentle (1980) and Monahan (2001), typically describe the implementation of matrix and linear-algebra computations on digital computers.

[13] It is also possible to find the Cholesky factor of a symmetric positive-semidefinite matrix, but then one or more diagonal elements of $\mathbf{U}$ will be 0, along with the other entries in the corresponding row. Notice, in addition, that in solving for the diagonal entries of $\mathbf{U}$, we must take positive square roots.

CHAPTER 2. AN INTRODUCTION TO CALCULUS

What is now called *calculus* deals with two basic types of problems: finding the slopes of tangent lines to curves (*differential calculus*) and evaluating areas under curves (*integral calculus*). In the 17th century, the English physicist and mathematician Sir Isaac Newton (1643–1727) and the German philosopher and mathematician Gottfried Wilhelm Leibniz (1646–1716) independently demonstrated the relationship between these two kinds of problems, consolidating and extending previous work in mathematics dating to the classical period. Newton and Leibniz are generally acknowledged as the cofounders of calculus.[1] In the 19th century, the great French mathematician Augustin Louis Cauchy (1789–1857), among others, employed the concept of the limit of a function to provide a rigorous logical foundation for calculus.

After a review of some elementary mathematics—numbers, equations of lines and planes, polynomial functions, logarithms, exponentials, and basic trigonometric functions—I will briefly take up the following seminal topics in calculus, emphasizing basic concepts: Section 2.2, limits of functions; Section 2.3, the derivative of a function; Section 2.4, the application of derivatives to optimization problems; Section 2.5, partial derivatives of functions of several variables, constrained optimization, and differential calculus in matrix form; Section 2.6, Taylor series expansions and approximations; and Section 2.7, the essential ideas of integral calculus.

Although a thorough and rigorous treatment is well beyond the scope of this brief book, one can get a lot of mileage out of an intuitive grounding in the basic ideas of calculus.

2.1 Review

2.1.1 Numbers

The definition of various sets of numbers is a relatively deep topic in mathematics, but the following rough distinctions will be sufficient for our purposes:

[1] Newton's claim that Leibniz had appropriated his work touched off one of the most famous priority disputes in the history of science.

- The *natural numbers* include 0 and the positive whole numbers: $0, 1, 2, 3, \ldots$.[2]
- The *integers* include all negative and positive whole numbers and 0:

$$\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots$$

- The *rational numbers* consist of all numbers that can be written as ratios of two integers, $\frac{n}{m}$ with $m \neq 0$, including all of the integers and nonintegers such as $-\frac{1}{2}$ and $\frac{123}{4}$.
- The *real numbers* include all of the rational numbers along with the *irrational numbers*, such as $\sqrt{2} \approx 1.41421$ and the mathematical constants $\pi \approx 3.14159$ and $e \approx 2.71828$, which cannot be written precisely as the ratio of two integers. The real numbers can be mapped into distances along a continuous line from $-\infty$ to $+\infty$.
- The *complex numbers* are of the form $a + bi$, where a and b are real numbers, and where $i \equiv \sqrt{-1}$. The complex numbers can be thought of as points in a plane: The real component of the number a gives the horizontal coordinate of the point, and the coefficient b of the "imaginary" component bi gives the vertical coordinate. The complex numbers include the real numbers (for which $b = 0$).

2.1.2 Lines and Planes

A *straight line* has the equation

$$y = a + bx$$

where a and b are constants. The constant a is the *y-intercept* of the line, that is, the value of y associated with $x = 0$; and b is the *slope* of the line, that is the change in y when x is increased by 1: See Figure 2.1, which shows straight lines in the two-dimensional coordinate space with axes x and y; in each case, the line extends infinitely to the left and right beyond the line-segment shown in the graph. When the slope is positive, $b > 0$, the line runs from lower left to upper right; when the slope is negative, $b < 0$, the line runs from upper left to lower right; and when $b = 0$, the line is horizontal.

Similarly, the *linear equation*

$$y = a + b_1 x_1 + b_2 x_2$$

represents a flat *plane* in the three-dimensional space with axes x_1, x_2, and y, as illustrated in the three-dimensional graph in Figure 2.2; the

[2] In some areas of mathematics, the natural numbers include only the positive whole numbers.

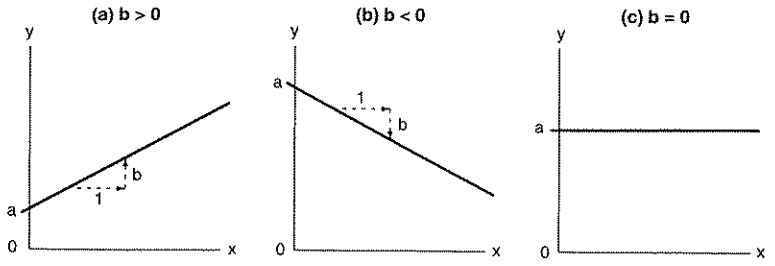

Figure 2.1 The graph of a straight line, $y = a + bx$, for (a) $b > 0$, (b) $b < 0$, and (c) $b = 0$.

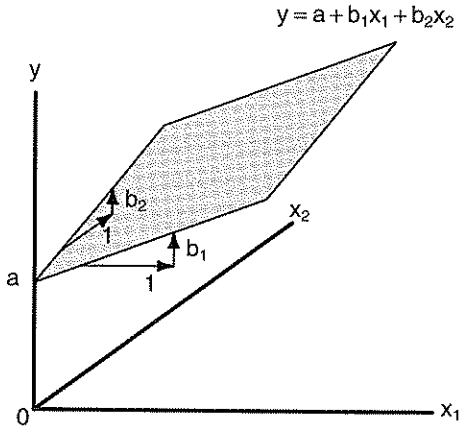

Figure 2.2 The equation of a plane, $y = a + b_1x_1 + b_2x_2$. Here, both slopes, b_1 and b_2, are positive.

axes are at right angles to each other, so think of the x_2 axis as extending directly into the page. The plane extends infinitely in all directions beyond the lines on its surface shown in the graph. The intercept of the plane, a, is the value of y when both x_1 and x_2 are 0; b_1 represents the slope of the plane in the direction of x_1 for a fixed value of x_2; and b_2 represents the slope of the plane in the direction of x_2 for a fixed value of x_1.

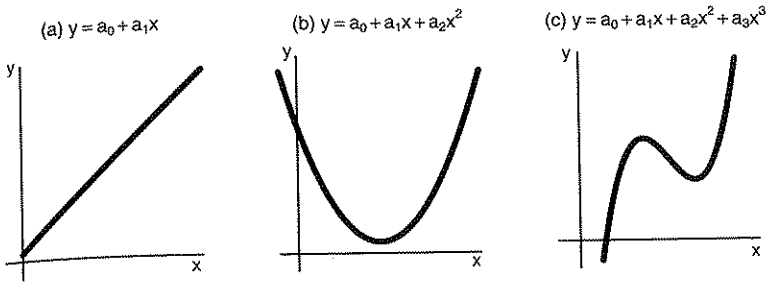

Figure 2.3 "Typical" first-order (linear), second-order (quadratic), and third-order (cubic) polynomials.

The equation of a straight line can be written in other forms, including

$$cx + dy = e$$

which can be transformed into slope-intercept form as

$$y = \frac{e}{d} - \frac{c}{d}x$$

Likewise, the equation

$$c_1x_1 + c_2x_2 + dy = e$$

represents a plane,

$$y = \frac{e}{d} - \frac{c_1}{d}x_1 - \frac{c_2}{d}x_2$$

2.1.3 Polynomials

Polynomials are functions of the form

$$y = a_0 + a_1x + a_2x^2 + \cdots + a_px^p$$

where $a_0, a_1, a_2, \ldots, a_p$ are constants, some of which (with the exception of a_p) may be 0. The largest exponent, p, is called the *order* of the polynomial. In particular, and as illustrated in Figure 2.3, a first-order polynomial is a straight line,

$$y = a_0 + a_1x$$

a second-order polynomial is a *quadratic equation*,

$$y = a_0 + a_1 x + a_2 x^2$$

and a third-order polynomial is a *cubic equation*,

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3$$

A polynomial equation of order p can have up to $p - 1$ "bends" in it, such as the single bend (change of direction) in the quadratic function in Figure 2.3(b) and the two bends in the cubic in Figure 2.3(c).

2.1.4 Logarithms and Exponentials

Logarithms ("*logs*") are exponents: The expression

$$\log_b x = y$$

which is read as "the log of x to the base b is y," means that

$$x = b^y$$

where $b > 0$ and $b \neq 1$. Thus, for example,

$$\log_{10} 10 = 1 \text{ because } 10^1 = 10$$
$$\log_{10} 100 = 2 \text{ because } 10^2 = 100$$
$$\log_{10} 1 = 0 \text{ because } 10^0 = 1$$
$$\log_{10} 0.1 = -1 \text{ because } 10^{-1} = 0.1$$

and, similarly,

$$\log_2 2 = 1 \text{ because } 2^1 = 2$$
$$\log_2 4 = 2 \text{ because } 2^2 = 4$$
$$\log_2 1 = 0 \text{ because } 2^0 = 1$$
$$\log_2 \frac{1}{4} = -2 \text{ because } 2^{-2} = \frac{1}{4}$$

Indeed, the log of 1 to any base is 0, because $b^0 = 1$ for any number $b \neq 0$. Logs are defined only for positive numbers x. The most commonly used base for logarithms in mathematics is the base $e \approx 2.718$; logs to the

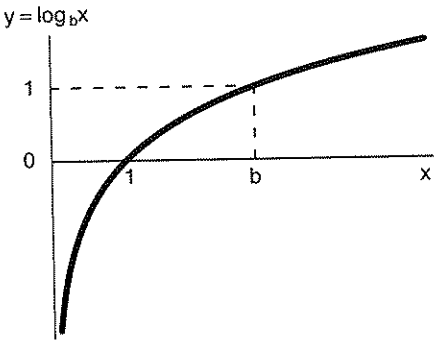

Figure 2.4 Graph of the log function $y = \log_b x$.

base e are called *natural logs*.[3] (For a justification of this terminology, see Section 2.3.4.)

A "typical" log function is graphed in Figure 2.4. As the graph implies, log functions have the same basic shape regardless of the base, and converting from one base, say b, to another, say a, simply involves multiplication by a constant:

$$\log_a x = \log_a b \times \log_b x$$

For example,

$$\log_{10} 1000 = 3 = \log_{10} 2 \times \log_2 1000 \approx 0.301030 \times 9.965784$$

Logs inherit their properties from the properties of exponents: Because $b^{x_1} b^{x_2} = b^{x_1 + x_2}$, it follows that

$$\log(x_1 x_2) = \log x_1 + \log x_2$$

Similarly, because $b^{x_1} / b^{x_2} = b^{x_1 - x_2}$,

$$\log\left(\frac{x_1}{x_2}\right) = \log x_1 - \log x_2$$

[3] Although I prefer always to show the base of the log function explicitly, as in $\log_{10}$ or $\log_e$ (unless the base is irrelevant, in which case log will do), many authors use unsubscripted log or ln to represent natural logs.

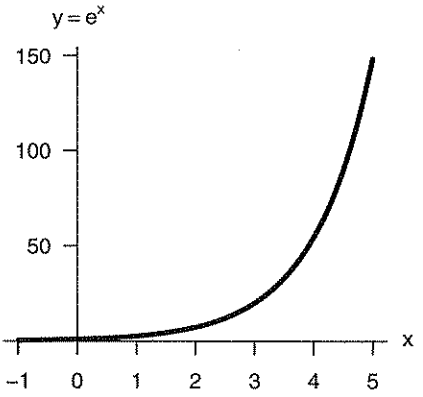

Figure 2.5 Graph of the exponential function $y = e^x$.

and because $b^{ax} = (b^x)^a$,

$$\log(x^a) = a \log x$$

At one time, the conversion of multiplication into addition, division into subtraction, and exponentiation into multiplication simplified laborious computations. Although this motivation has faded, logs still play a prominent role in mathematics and statistics.

An *exponential function* is a function of the form

$$y = a^x$$

where a is a constant. The most common exponential, $y = \exp(x) = e^x$, is graphed in Figure 2.5. The log and exponential functions are inverses of each other, in the sense that $\log_a(a^x) = x$ and $a^{\log_a x} = x$.

2.1.5 Basic Trigonometric Functions

Figure 2.6 shows a *unit circle*—that is, a circle of radius 1 centered at the origin. The angle x produces a right triangle inscribed in the circle; notice that the angle is measured in a counterclockwise direction from the horizontal axis. The cosine of the angle x, denoted $\cos x$, is the signed length of the side of the triangle adjacent to the angle (i.e., "adjacent/hypotenuse,"

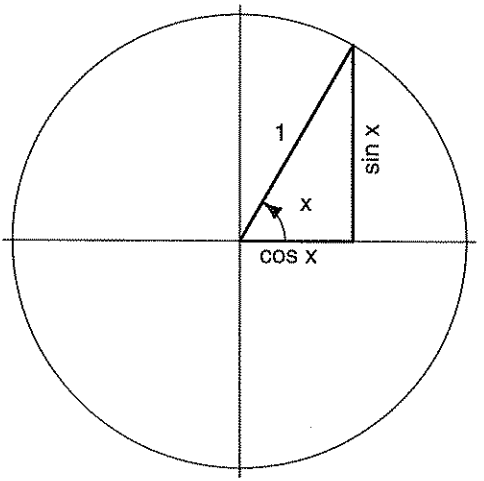

Figure 2.6 A unit circle, showing the angle x and its cosine and sine.

where the hypotenuse is 1 because it is a radius of the unit circle); the sine of the angle x, denoted $\sin x$, is the signed length of the side of the triangle opposite the angle (i.e., "opposite/hypotenuse"); and the tangent of x, $\tan x = \sin x / \cos x$, is the ratio of the signed length of the side opposite to the side adjacent to the right angle ("opposite"/"adjacent"). The cosine, sine, and tangent functions for angles between $-360°$ and $360°$ are shown in Figure 2.7; negative angles represent clockwise rotations. Notice that the tangent function approaches $\pm\infty$ at angles of $\pm 90°$ and $\pm 270°$, and that the sine and cosine functions have the same shape, with $\sin(x) = \cos(x - 90)$.

It is sometimes mathematically convenient to measure angles in *radians* rather than in degrees, with 2π radians corresponding to 360 degrees. The circumferance of the unit circle in Figure 2.6 is also 2π, and therefore the radian measure of an angle represents the length of the arc along the unit circle subtended by the angle.

2.2 Limits

Calculus deals with *functions* of the form $y = f(x)$. I will consider the case where both the *domain* (values of the *independent variable* x) and *range* (values of the *dependent variable* y) of the function are real numbers. The

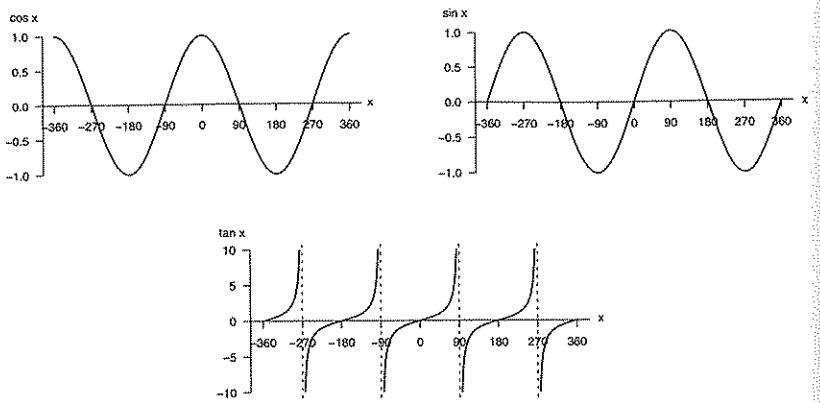

Figure 2.7 The cosine, sine, and tangent functions for angles between $x = -360°$ and $x = 360°$.

limit of a function concerns its behavior when x is near, but not necessarily equal to, a specific value. This is often a useful idea, especially when a function is undefined at a particular value of x.

2.2.1 The "Epsilon-Delta" Definition of a Limit

A function $y = f(x)$ has a limit L at $x = x_0$ (i.e., a particular value of x) if for any positive *tolerance* ε, no matter how small, there exists a positive number δ such that the distance between $f(x)$ and L is less than the tolerance as long as the distance between x and x_0 is smaller than δ—that is, as long as x is confined to a sufficiently small *neighborhood* of width 2δ around x_0. In symbols:

$$|f(x) - L| < \varepsilon$$

for all

$$0 < |x - x_0| < \delta$$

This possibly cryptic definition is clarified by Figure 2.8. Note that $f(x_0)$ need not equal L. Indeed, limits are often most useful when $f(x)$ does not exist at $x = x_0$. For L to be the limit of $f(x)$ at $x = x_0$, the function must approach this value as x approaches x_0 *both* from the left and from the right.

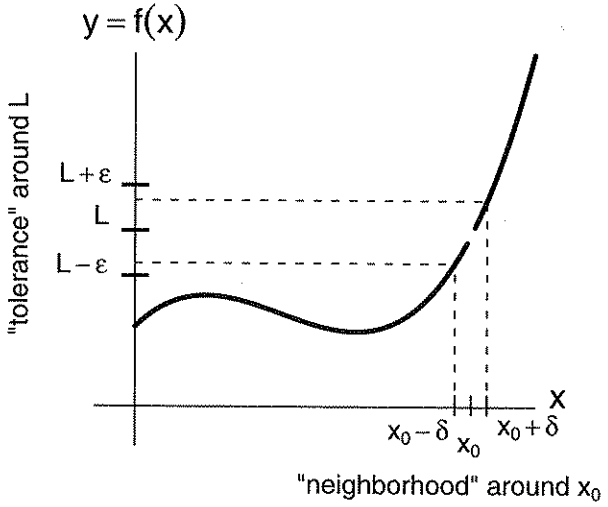

Figure 2.8 $\lim_{x \to x_0} f(x) = L$: The limit of the function $f(x)$ as x approaches x_0 is L. The gap in the curve above x_0 is meant to suggest that the function is undefined at $x = x_0$.

The following notation is used:

$$\lim_{x \to x_0} f(x) = L$$

We read this expression as "The limit of the function $f(x)$ as x approaches x_0 is L."

2.2.2 Finding a Limit: An Example

Find the limit of

$$y = f(x) = \frac{x^2 - 1}{x - 1}$$

at $x_0 = 1$:

Notice that $f(1) = \dfrac{1-1}{1-1} = \dfrac{0}{0}$ is undefined. Nevertheless, as long as x is not *exactly* equal to 1, even if it is very close to it, we can divide by $x - 1$:

$$y = \frac{x^2 - 1}{x - 1} = \frac{(x+1)(x-1)}{x-1} = x + 1$$

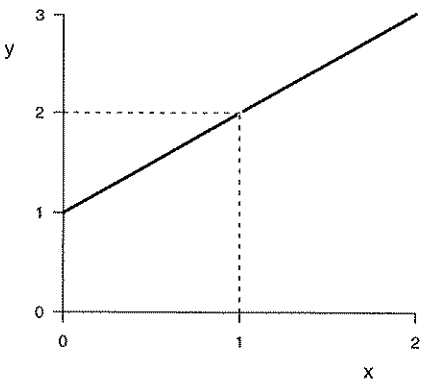

Figure 2.9 $\lim_{x\to 1} \frac{x^2-1}{x-1} = 2$, even though the function is undefined at $x = 1$.

Moreover, because $x_0 + 1 = 1 + 1 = 2$,

$$\lim_{x\to 1} \frac{x^2 - 1}{x - 1} = \lim_{x\to 1} (x + 1)$$
$$= 1 + 1 = 2$$

This limit is graphed in Figure 2.9.

2.2.3 Rules for Manipulating Limits

Suppose that we have two functions $f(x)$ and $g(x)$ of an independent variable x, and that each function has a limit at $x = x_0$:

$$\lim_{x\to x_0} f(x) = a$$
$$\lim_{x\to x_0} g(x) = b$$

Then the limits of functions composed from $f(x)$ and $g(x)$ by the arithmetic operations of addition, subtraction, multiplication, and division are straightforward:

$$\lim_{x\to x_0} \big[f(x) + g(x)\big] = a + b$$
$$\lim_{x\to x_0} \big[f(x) - g(x)\big] = a - b$$

$$\lim_{x \to x_0} \left[f(x)g(x) \right] = ab$$

$$\lim_{x \to x_0} \left[f(x)/g(x) \right] = a/b$$

The last result holds as long as the denominator $b \neq 0$.

Similarly, if c and n are constants and $\lim_{x \to x_0} f(x) = a$, then

$$\lim_{x \to x_0} c = c$$

$$\lim_{x \to x_0} \left[cf(x) \right] = ca$$

$$\lim_{x \to x_0} \left\{ \left[f(x) \right]^n \right\} = a^n$$

Finally, it is (I hope) obvious that

$$\lim_{x \to x_0} x = x_0$$

2.3 The Derivative of a Function

Now consider a function $y = f(x)$ evaluated at two values of x:

$$\begin{aligned} &\text{at } x_1: \quad y_1 = f(x_1) \\ &\text{at } x_2: \quad y_2 = f(x_2) \end{aligned}$$

The *difference quotient* is defined as the change in y divided by the change in x, as we move from the point (x_1, y_1) to the point (x_2, y_2):

$$\frac{y_2 - y_1}{x_2 - x_1} = \frac{\Delta y}{\Delta x} = \frac{f(x_2) - f(x_1)}{x_2 - x_1}$$

where Δ ("Delta") is a short-hand denoting "change." As illustrated in Figure 2.10, the difference quotient is the slope of the *secant line* connecting the points (x_1, y_1) and (x_2, y_2).

The *derivative* of the function $f(x)$ at $x = x_1$ (so named because it is *derived* from the original function) is the limit of the difference quotient $\Delta y / \Delta x$ as x_2 approaches x_1 (i.e., as $\Delta x \to 0$):

$$\begin{aligned} \frac{dy}{dx} &= \lim_{x_2 \to x_1} \frac{f(x_2) - f(x_1)}{x_2 - x_1} \\ &= \lim_{\Delta x \to 0} \frac{f(x_1 + \Delta x) - f(x_1)}{\Delta x} \\ &= \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} \end{aligned}$$

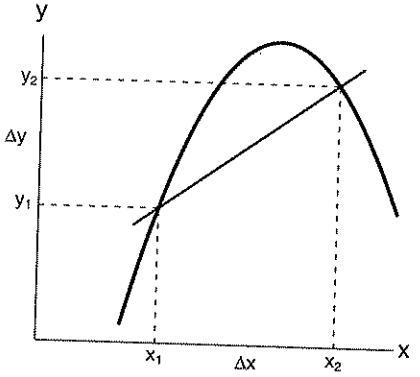

Figure 2.10 The difference quotient $\Delta y/\Delta x$ is the slope of the secant line connecting (x_1, y_1) and (x_2, y_2).

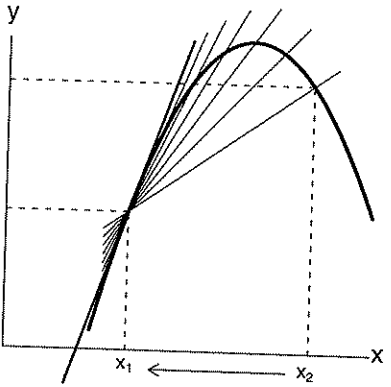

Figure 2.11 The derivative is the slope of the tangent line at $f(x_1)$. As $x_2 \to x_1$, the secant line approaches the tangent line.

The derivative is therefore the slope of the *tangent line* to the curve $f(x)$ at $x = x_1$, as shown in Figure 2.11.

The following alternative notation is often used for the derivative:

$$\frac{dy}{dx} = \frac{df(x)}{dx} = f'(x)$$

The last form, $f'(x)$, emphasizes that the derivative is itself a function of x, but the notation employing the *differentials* dy and dx, which may be thought of as infinitesimally small values that are nevertheless nonzero, can be productive: In many circumstances the differentials can be manipulated as if they were numbers. (See, e.g., the "chain rule" for differentiation, introduced in Section 2.3.3.) The operation of finding the derivative of a function is called *differentiation*.

2.3.1 The Derivative as the Limit of the Difference Quotient: An Example

Given the function $y = f(x) = x^2$, find the derivative $f'(x)$ for any value of x:

Applying the definition of the derivative as the limit of the difference quotient,

$$\begin{aligned} f'(x) &= \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x} \\ &= \lim_{\Delta x \to 0} \frac{(x + \Delta x)^2 - x^2}{\Delta x} \\ &= \lim_{\Delta x \to 0} \frac{x^2 + 2x\Delta x + (\Delta x)^2 - x^2}{\Delta x} \\ &= \lim_{\Delta x \to 0} \frac{2x\Delta x + (\Delta x)^2}{\Delta x} \\ &= \lim_{\Delta x \to 0} (2x + \Delta x) \\ &= \lim_{\Delta x \to 0} 2x + \lim_{\Delta x \to 0} \Delta x \\ &= 2x + 0 = 2x \end{aligned}$$

Notice that division by Δx is justified here, because although Δx approaches 0 in the limit, it never is exactly equal to 0. For example, the slope of the curve $y = f(x) = x^2$ at $x = 3$ is $f'(x) = 2x = 2 \times 3 = 6$.

2.3.2 Derivatives of Powers

More generally, by similar reasoning, the derivative of

$$y = f(x) = ax^n$$

is

$$\frac{dy}{dx} = nax^{n-1}$$

For example, the derivative of the function

$$y = 3x^6$$

is

$$\frac{dy}{dx} = 6 \times 3x^{6-1} = 18x^5$$

Moreover, this rule applies as well to negative powers and to fractional powers. For example, the derivative of the function

$$y = \frac{1}{4x^3} = \frac{1}{4}x^{-3}$$

is

$$\frac{dy}{dx} = -3 \times \frac{1}{4}x^{-3-1} = -\frac{3}{4}x^{-4} = -\frac{3}{4x^4}$$

and the derivative of the function

$$y = \sqrt{x} = x^{\frac{1}{2}}$$

is

$$\frac{dy}{dx} = \tfrac{1}{2}x^{\frac{1}{2}-1} = \tfrac{1}{2}x^{-\frac{1}{2}} = \frac{1}{2\sqrt{x}}$$

2.3.3 Rules for Manipulating Derivatives

Suppose that a function is the sum of two other functions:

$$h(x) = f(x) + g(x)$$

The *addition rule* for derivatives follows from the addition rule for limits:

$$h'(x) = f'(x) + g'(x)$$

For example,

$$y = 2x^2 + 3x + 4$$
$$\frac{dy}{dx} = 4x + 3 + 0 = 4x + 3$$

Notice that the derivative of a constant—the constant 4 in the last example—is 0, because the constant can be expressed as

$$y = f(x) = 4 = 4x^0$$

This result makes sense geometrically: A constant is represented as a horizontal line in the $\{x, y\}$ plane, and a horizontal line has a slope of 0.

The addition rule, therefore, along with the result that $\frac{d}{dx}ax^n = nax^{n-1}$, serves to differentiate any polynomial function.

Multiplication and division are more complex. The *multiplication rule* for derivatives:

$$h(x) = f(x)g(x)$$
$$h'(x) = f(x)g'(x) + f'(x)g(x)$$

The *division rule* for derivatives:

$$h(x) = f(x)/g(x)$$
$$h'(x) = \frac{g(x)f'(x) - g'(x)f(x)}{\left[g(x)\right]^2}$$

For example, the derivative of the function

$$y = (x^2 + 1)(2x^3 - 3x)$$

is

$$\frac{dy}{dx} = (x^2 + 1)(6x^2 - 3) + 2x(2x^3 - 3x)$$

and the derivative of the function

$$y = \frac{x}{x^2 - 3x + 5}$$

is

$$\frac{dy}{dx} = \frac{x^2 - 3x + 5 - (2x - 3)x}{(x^2 - 3x + 5)^2} = \frac{-x^2 + 5}{(x^2 - 3x + 5)^2}$$

The *chain rule*: If $y = f(z)$ and $z = g(x)$, then y is indirectly a function of x:

$$y = f[g(x)] = h(x)$$

The derivative of y with respect to x is

$$h'(x) = \frac{dy}{dx} = \frac{dy}{dz} \times \frac{dz}{dx}$$

as if the differential dz in the numerator and the denominator can be cancelled.[4]

For example, given the function

$$y = (x^2 + 3x + 6)^5$$

find the derivative dy/dx of y with respect to x:

This problem could be solved by expanding the power—that is, by multiplying the expression in parentheses by itself five times—but that would be tedious in the extreme. It is much easier to find the derivative by using the chain rule, introducing a new variable, z, to represent the expression inside the parentheses. Let

$$z = g(x) = x^2 + 3x + 6$$

Then

$$y = f(z) = z^5$$

Differentiating y with respect to z, and z with respect to x, produces

$$\frac{dy}{dz} = 5z^4$$

$$\frac{dz}{dx} = 2x + 3$$

[4]The differentials are not ordinary numbers, so thinking of the chain rule as simultaneously dividing and multiplying by the differential dz is a heuristic device, illustrating how the notation for the derivative using differentials proves to be productive.

Applying the chain rule,

$$\frac{dy}{dx} = \frac{dy}{dz} \times \frac{dz}{dx}$$
$$= 5z^4(2x+3)$$

Finally, substituting for z,

$$\frac{dy}{dx} = 5(x^2+3x+6)^4(2x+3)$$

The use of the chain rule in this example is typical, introducing an "artificial" variable (z) to simplify the structure of the problem.

2.3.4 Derivatives of Logs and Exponentials

Logarithms and exponentials often occur in statistical applications, and so it is useful to know how to differentiate these functions.

The derivative of the log function $y = \log_e(x)$ is

$$\frac{d\log_e x}{dx} = \frac{1}{x} = x^{-1}$$

Recall that $\log_e$ is the *natural-log* function, that is, log to the base $e \approx 2.718$. Indeed, the simplicity of its derivative is one of the reasons that it is "natural" to use the base e for the natural logs.

The derivative of the exponential function $y = e^x$ is

$$\frac{de^x}{dx} = e^x$$

The derivative of the exponential function $y = a^x$ for any constant a (i.e., not necessarily e) is

$$\frac{da^x}{dx} = a^x \log_e a$$

2.3.5 Derivatives of the Basic Trigonometric Functions

The derivatives of the basic trigometric functions are as follows, with the angle x measured in radians:

$$\frac{d\cos x}{dx} = -\sin x$$
$$\frac{d\sin x}{dx} = \cos x$$

$$\frac{d \tan x}{dx} = \frac{1}{\cos^2 x}$$

$$\text{for } x \neq \pm\frac{\pi}{2}, \pm\frac{3\pi}{2}, \text{ etc. [i.e., where } \cos(x) \neq 0]$$

Note that $\cos^2 x \equiv (\cos x)^2$.

2.3.6 Second-Order and Higher-Order Derivatives

Because derivatives are themselves functions, they can be differentiated. The *second derivative* of the function $y = f(x)$ is therefore defined as

$$f''(x) = \frac{d^2y}{dx^2} = \frac{df'(x)}{dx}$$

Notice the alternative notation.

Likewise, the *third derivative* of the function $y = f(x)$ is the derivative of the second derivative,

$$f'''(x) = \frac{d^3y}{dx^3} = \frac{df''(x)}{dx}$$

and so on for *higher-order derivatives*.

For example, the function

$$y = f(x) = 5x^4 + 3x^2 + 6$$

has the derivatives

$$f'(x) = 20x^3 + 6x$$
$$f''(x) = 60x^2 + 6$$
$$f'''(x) = 120x$$
$$f''''(x) = 120$$
$$f'''''(x) = 0$$

All derivatives beyond the fifth-order are also 0.

2.4 Optimization

An important application of derivatives, both in statistics and more generally, is to *maximization* and *minimization* problems: that is, finding maximum and

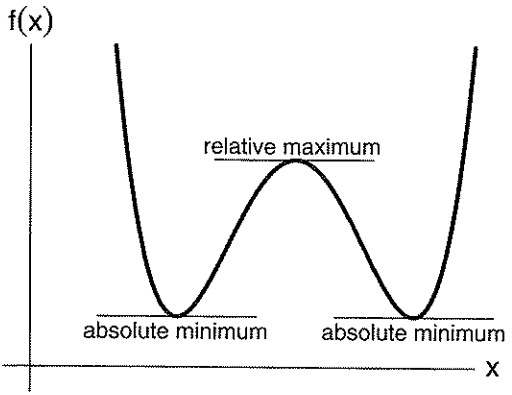

Figure 2.12 The derivative (i.e., the slope) of the function is 0 where the function $f(x)$ is at a minimum or maximum.

minimum values of functions (e.g., *maximum*-likelihood estimation; *least*-squares estimation). Such problems are collectively called *optimization*.

As illustrated in Figure 2.12, when a function is at a *relative (local) maximum* or *relative minimum* (i.e., a value higher than or lower than surrounding values) or at an *absolute* or *global maximum* or *minimum* (a value at least as high or low as all other values of the function), the tangent line to the function is flat, and hence the function has a derivative of 0 at that point. A function can also have a 0 derivative, however, at a point that is neither a minimum nor a maximum, such as at a *point of inflection*—that is, a point where the direction of curvature changes, as in Figure 2.13. Points at which the derivative is 0 are called *stationary points*.

To distinguish among the three cases—minimum, maximum, or neither—we can appeal to the value of the second derivative (see Figure 2.14).

- At a *minimum*, the first derivative $f'(x)$ is changing from negative, through 0, to positive—that is, the first derivative is *increasing*, and therefore the second derivative $f''(x)$ is *positive*: The second derivative indicates change in the first derivative just as the first derivative indicates change in the original function.
- At a *maximum*, the first derivative $f'(x)$ is changing from positive, through 0, to negative—the first derivative is *decreasing*, and therefore the second derivative $f''(x)$ is *negative*.
- At a point of inflection, $f''(x) = 0$.

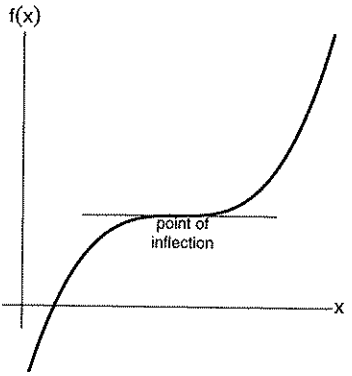

Figure 2.13 The derivative is also 0 at a point of inflection in $f(x)$.

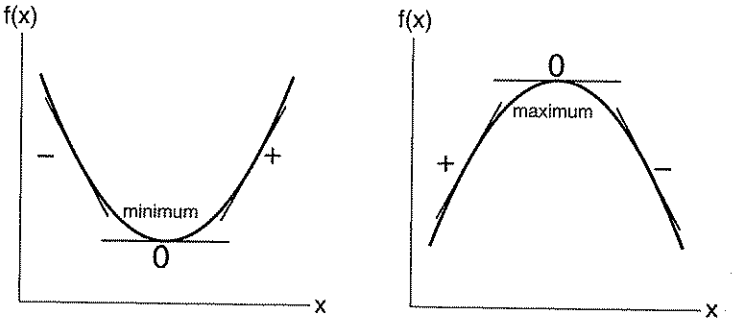

Figure 2.14 The first derivative (the slope of the function) is increasing where the function $f(x)$ is at a minimum and decreasing at a maximum.

The relationships among the original function, the first derivative, and the second derivative are illustrated in Figure 2.15: The first derivative dy/dx is 0 at the two minima and at the (relative) maximum of $f(x)$; the second derivative d^2y/dx^2 is positive at the two minima, and negative at the maximum of $f(x)$.

2.4.1 Optimization: An Example

Find the *extrema* (minima and maxima) of the function

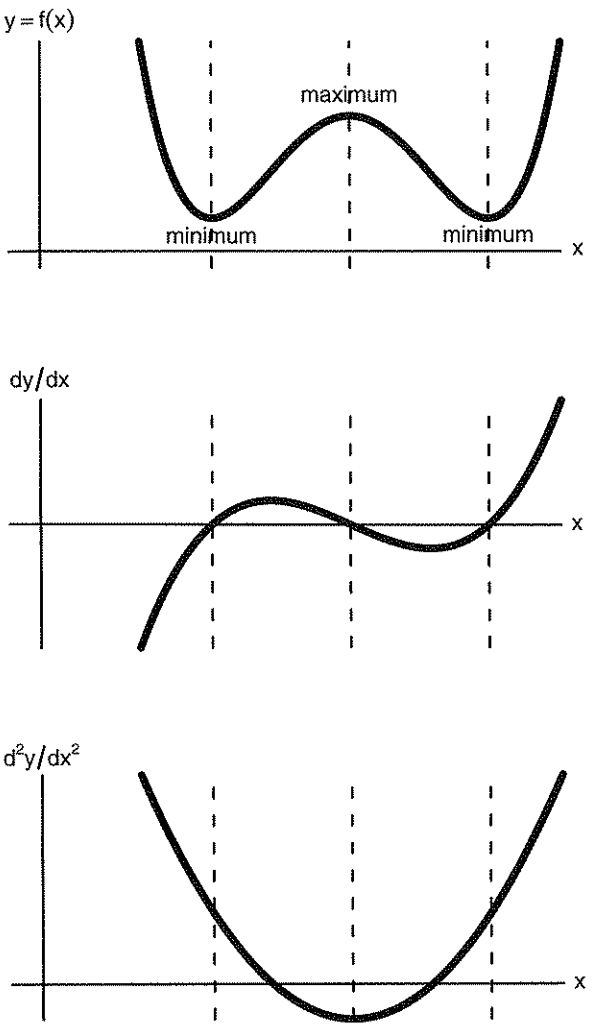

Figure 2.15 An example of a function and its first and second derivatives.

$$f(x) = 2x^3 - 9x^2 + 12x + 6$$

The function is shown in Figure 2.16. By the way, locating stationary points and determining whether they are minima or maxima (or neither) is helpful in graphing functions.

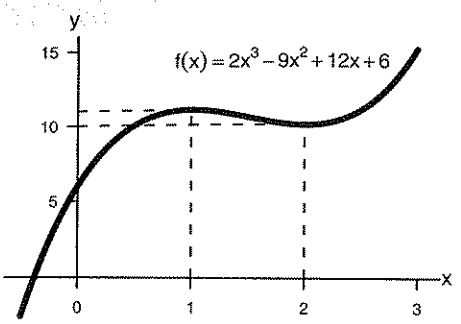

Figure 2.16 Finding the extrema of the function $y = f(x) = 2x^3 - 9x^2 + 12x + 6$.

The first and second derivatives of the function are

$$f'(x) = 6x^2 - 18x + 12$$
$$f''(x) = 12x - 18$$

Setting the first derivative to 0, and solving for the values of x that satisfy the resulting equation, produces the following results:

$$6x^2 - 18x + 12 = 0$$
$$x^2 - 3x + 2 = 0$$
$$(x - 2)(x - 1) = 0$$

The two roots, at which $f'(x)$ is 0, are therefore $x = 2$ and $x = 1$.

- For $x = 2$,

$$f(2) = 2 \times 2^3 - 9 \times 2^2 + 12 \times 2 + 6 = 10$$
$$f'(2) = 6 \times 2^2 - 18 \times 2 + 12 = 0\checkmark$$
$$f''(2) = 12 \times 2 - 18 = 6$$

  Because $f''(2)$ is *positive*, the point (2, 10) represents a (relative) *minimum*.
- Likewise, for $x = 1$,

$$f(1) = 2 \times 1^3 - 9 \times 1^2 + 12 \times 1 + 6 = 11$$
$$f'(1) = 6 \times 1^2 - 18 \times 1 + 12 = 0\checkmark$$
$$f''(1) = 12 \times 1 - 18 = -6$$

  Because $f''(1)$ is *negative*, the point (1, 11) represents a (relative) *maximum*.

2.5 Multivariable and Matrix Differential Calculus

Multivariable differential calculus—the topic of this section—finds frequent application in statistics. The essential ideas of multivariable calculus are straightforward extensions of calculus of a single independent variable, but the topic is frequently omitted from introductory treatments of calculus.

2.5.1 Partial Derivatives

Consider a function $y = f(x_1, x_2, \ldots, x_n)$ of several independent variables. The *partial derivative* of y with respect to a particular x_i is the derivative of $f(x_1, x_2, \ldots, x_n)$ treating the other xs as constants. To distinguish it from the ordinary derivative dy/dx, the standard notation for the partial derivative uses "curly ds" in place of ds: $\partial y/\partial x_i$.

For example, for the function

$$y = f(x_1, x_2) = x_1^2 + 3x_1x_2^2 + x_2^3 + 6$$

the partial derivatives with respect to x_1 and x_2 are

$$\frac{\partial y}{\partial x_1} = 2x_1 + 3x_2^2 + 0 + 0 = 2x_1 + 3x_2^2$$
$$\frac{\partial y}{\partial x_2} = 0 + 6x_1x_2 + 3x_2^2 + 0 = 6x_1x_2 + 3x_2^2$$

The "trick" in partial differentiation with respect to x_i is to remember to treat all of the other xs as constants (i.e., literally to hold other xs constant). Thus, when we differentiate with respect to x_1, terms such as x_2^2 and x_2^3 are constants.

The partial derivative $\partial f(x_1, x_2, \ldots, x_n)/\partial x_1$ gives the slope of the tangent hyperplane to the function $f(x_1, x_2, \ldots, x_n)$ in the direction of x_1.[5] For example, the tangent plane to the function

$$f(x_1, x_2) = x_1^2 + x_1x_2 + x_2^2 + 10$$

above the pair of values $x_1 = 1, x_2 = 2$ is shown in Figure 2.17.

At a local or global minimum or maximum, the slope of the tangent hyperplane is 0 in all directions. Consequently, to minimize or maximize

[5] A *hyperplane* is the generalization of a linear (i.e., flat) surface to a space of more than three dimensions. The dimension of the hyperplane is one less than that of the enclosing space, just as a plane is a two-dimensional object embedded in a three-dimensional space.

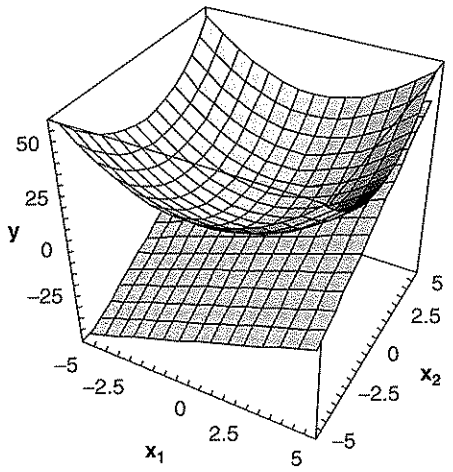

Figure 2.17 The function $y = f(x_1, x_2) = x_1^2 + x_1x_2 + x_2^2 + 10$, showing the tangent plane at $x_1 = 1, x_2 = 2$.

a function of several variables, we have to differentiate the function with respect to each variable, set the partial derivatives to 0, and solve the resulting set of simultaneous equations. I will explain in Section 2.5.3 how to distinguish maxima from minima.

Let us, for example, find the values of x_1 and x_2 that minimize the function

$$y = f(x_1, x_2) = x_1^2 + x_1x_2 + x_2^2 + 10$$

Differentiating,

$$\frac{\partial y}{\partial x_1} = 2x_1 + x_2$$
$$\frac{\partial y}{\partial x_2} = x_1 + 2x_2$$

Setting these partial derivatives to 0 produces the unique solution $x_1 = 0$, $x_2 = 0$. In this case, the solution is particularly simple because the partial derivatives are linear functions of x_1 and x_2. The value of the function at its minimum is

$$y = 0^2 + (0 \times 0) + 0^2 + 10 = 10$$

The slopes of the tangent plane above the pair of values $x_1 = 1$, $x_2 = 2$, illustrated in Figure 2.17, are

$$\frac{\partial y}{\partial x_1} = 2(1) + 2 = 4$$
$$\frac{\partial y}{\partial x_2} = 1 + 2(2) = 5$$

2.5.2 Lagrange Multipliers for Constrained Optimization

The method of Lagrange multipliers (named after the 18th-century French mathematician Joseph-Louis Lagrange) permits us to optimize a function of the form $y = f(x_1, x_2, \ldots, x_n)$ subject to a constraint of the form $g(x_1, x_2, \ldots, x_n) = 0$. The method, in effect, incorporates the constraint into the set of partial derivatives.

Here is a simple example: Minimize

$$y = f(x_1, x_2) = x_1^2 + x_2^2$$

subject to the restriction that $x_1 + x_2 = 1$. (In the absence of this restriction, it is obvious that $x_1 = x_2 = 0$ minimizes the function.) To solve this constrained minimization problem:

1. Rewrite the constraint in the required form, $g(x_1, x_2, \ldots, x_n) = 0$. That is, $x_1 + x_2 - 1 = 0$.
2. Construct a new function incorporating the constraint. In the general case, this function takes the form[6]

   $$h(x_1, x_2, \ldots, x_n, \lambda) \equiv f(x_1, x_2, \ldots, x_n) - \lambda \times g(x_1, x_2, \ldots, x_n)$$

   The new independent variable λ is called a *Lagrange multiplier*. For the example,

   $$h(x_1, x_2, \lambda) \equiv x_1^2 + x_2^2 - \lambda(x_1 + x_2 - 1)$$

3. Find the values of $x_1, x_2, \ldots, x_n$ that (along with λ) optimize the function $h(x_1, x_2, \ldots, x_n, \lambda)$. That is, differentiate $h(x_1, x_2, \ldots, x_n, \lambda)$ with respect to each of $x_1, x_2, \ldots, x_n$ and λ; set the $n + 1$ partial derivatives to 0; and solve

[6] Some authors prefer to add, rather than subtract, the constraint,

$$h(x_1, x_2, \ldots, x_n, \lambda) \equiv f(x_1, x_2, \ldots, x_n) + \lambda \times g(x_1, x_2, \ldots, x_n)$$

but, except for a change in the sign of λ, the two approaches are equivalent.

the resulting system of simultaneous equations for $x_1, x_2, \ldots, x_n$ and λ. For the example,

$$\frac{\partial h(x_1, x_2, \lambda)}{\partial x_1} = 2x_1 - \lambda$$
$$\frac{\partial h(x_1, x_2, \lambda)}{\partial x_2} = 2x_2 - \lambda$$
$$\frac{\partial h(x_1, x_2, \lambda)}{\partial \lambda} = -x_1 - x_2 + 1$$

Notice that the partial derivative with respect to λ, when equated to 0, reproduces the constraint $x_1 + x_2 = 1$. Consequently, whatever solutions satisfy the equations produced by setting the partial derivatives to 0, necessarily satisfy the constraint. In this case, there is only one solution: $x_1 = x_2 = 0.5$ (and $\lambda = 1$).

The method of Lagrange multipliers easily extends to handle several restrictions, by introducing a separate Lagrange multiplier for each restriction.

2.5.3 Differential Calculus in Matrix Form

The function $y = f(x_1, x_2, \ldots, x_n)$ of the independent variables $x_1, x_2, \ldots, x_n$ can be written as the function $y = f(\mathbf{x})$ of the vector $\mathbf{x} = [x_1, x_2, \ldots, x_n]'$. The *vector partial derivative* (or the *gradient*) of y with respect to $\mathbf{x}$ is defined as the column vector of partial derivatives of y with respect to each of the entries of $\mathbf{x}$:

$$\frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} \\ \frac{\partial y}{\partial x_2} \\ \vdots \\ \frac{\partial y}{\partial x_n} \end{bmatrix}$$

If, therefore, y is a linear function of $\mathbf{x}$,

$$y = \underset{(1\times n)}{\mathbf{a}'} \underset{(n\times 1)}{\mathbf{x}} = a_1x_1 + a_2x_2 + \cdots + a_nx_n$$

then $\partial y/\partial x_i = a_i$, and $\partial y/\partial \mathbf{x} = \mathbf{a}$. For example, for

$$y = x_1 + 3x_2 - 5x_3$$
$$= [1, 3, -5] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

the vector partial derivative is

$$\frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} 1 \\ 3 \\ -5 \end{bmatrix}$$

Alternatively, suppose that y is a quadratic form in $\mathbf{x}$ (see Section 1.6),

$$y = \underset{(1\times n)}{\mathbf{x}'} \underset{(n\times n)}{\mathbf{A}} \underset{(n\times 1)}{\mathbf{x}}$$

where the matrix $\mathbf{A}$ is symmetric. Expanding the matrix product gives us

$$y = a_{11}x_1^2 + a_{22}x_2^2 + \cdots + a_{nn}x_n^2 + 2a_{12}x_1x_2 + \cdots + 2a_{1n}x_1x_n + \cdots + 2a_{n-1,n}x_{n-1}x_n$$

and, thus,

$$\frac{\partial y}{\partial x_i} = 2(a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n) = 2\mathbf{a}_i'\mathbf{x}$$

where $\mathbf{a}_i'$ represents the ith row of $\mathbf{A}$. Placing these partial derivatives in a vector produces $\partial y/\partial \mathbf{x} = 2\mathbf{A}\mathbf{x}$. The vector partial derivatives of linear and quadratic functions are strikingly similar to the analogous scalar derivatives of functions of one variable: $d(ax)/dx = a$ and $d(ax^2)/dx = 2ax$.

For example, for

$$\begin{aligned} y &= [x_1, x_2]\begin{bmatrix} 2 & 3 \\ 3 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= 2x_1^2 + 3x_1x_2 + 3x_2x_1 + x_2^2 \\ &= 2x_1^2 + 6x_1x_2 + x_2^2 \end{aligned}$$

the partial derivatives are

$$\begin{aligned} \frac{\partial y}{\partial x_1} &= 4x_1 + 6x_2 \\ \frac{\partial y}{\partial x_2} &= 6x_1 + 2x_2 \end{aligned}$$

and the vector partial derivative is

$$\begin{aligned} \frac{\partial y}{\partial \mathbf{x}} &= \begin{bmatrix} 4x_1 + 6x_2 \\ 6x_1 + 2x_2 \end{bmatrix} \\ &= 2\begin{bmatrix} 2 & 3 \\ 3 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \checkmark \end{aligned}$$

The so-called *Hessian matrix* of second-order partial derivatives of the function $y = f(\mathbf{x})$ is defined in the following manner:

$$\frac{\partial^2 y}{\partial \mathbf{x}\, \partial \mathbf{x}'} = \begin{bmatrix} \frac{\partial^2 y}{\partial x_1^2} & \frac{\partial^2 y}{\partial x_1\, \partial x_2} & \cdots & \frac{\partial^2 y}{\partial x_1\, \partial x_n} \\ \frac{\partial^2 y}{\partial x_2\, \partial x_1} & \frac{\partial^2 y}{\partial x_2^2} & \cdots & \frac{\partial^2 y}{\partial x_2\, \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 y}{\partial x_n\, \partial x_1} & \frac{\partial^2 y}{\partial x_n\, \partial x_2} & \cdots & \frac{\partial^2 y}{\partial x_n^2} \end{bmatrix}$$

For instance, $\partial^2(\mathbf{x}'\mathbf{A}\mathbf{x})/\partial \mathbf{x}\, \partial \mathbf{x}' = 2\mathbf{A}$, for a symmetric matrix $\mathbf{A}$. The Hessian is named after the 19th-century German mathematician Ludwig Otto Hesse.

To minimize a function $y = f(\mathbf{x})$ of several variables, we can set the vector partial derivative to $\mathbf{0}$, $\partial y/\partial \mathbf{x} = \mathbf{0}$, and solve the resulting set of simultaneous equations for $\mathbf{x}$, obtaining a solution $\mathbf{x}^*$. This solution represents a (local) minimum of the function in question if the Hessian matrix evaluated at $\mathbf{x} = \mathbf{x}^*$ is positive definite. The solution represents a maximum if the Hessian is negative definite.[7] Again, there is a strong parallel with the scalar results for a single x: Recall that the second derivative d^2y/dx^2 is positive at a minimum and negative at a maximum.

I showed earlier that the function

$$y = f(x_1, x_2) = x_1^2 + x_1 x_2 + x_2^2 + 10$$

has a stationary point (i.e., a point at which the partial derivatives are 0) at $x_1 = x_2 = 0.5$. The second-order partial derivatives of this function are

$$\frac{\partial^2 y}{\partial x_1\, \partial x_2} = \frac{\partial^2 y}{\partial x_2\, \partial x_1} = 1$$
$$\frac{\partial^2 y}{\partial x_1^2} = \frac{\partial^2 y}{\partial x_2^2} = 2$$

The Hessian evaluated at $x_1 = x_2 = 0.5$ (or, indeed, at any point), is, therefore,

[7] The square matrix $\mathbf{H}$ (here, the Hessian) is *positive definite* if $\mathbf{x}'\mathbf{H}\mathbf{x} > 0$ for any nonzero vector $\mathbf{x}$. (See Section 1.6.) A positive-definite Hessian is a sufficient but not necessary condition for a minimum. Likewise, the square matrix $\mathbf{H}$ is *negative definite* if $\mathbf{x}'\mathbf{H}\mathbf{x} < 0$ for any nonzero vector $\mathbf{x}$; a negative-definite Hessian is a sufficient but not necessary condition for a maximum.

$$\begin{bmatrix} \dfrac{\partial^2 y}{\partial x_1^2} & \dfrac{\partial^2 y}{\partial x_1\,\partial x_2} \\ \dfrac{\partial^2 y}{\partial x_2\,\partial x_1} & \dfrac{\partial^2 y}{\partial x_2^2} \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

This matrix is clearly positive definite, verifying that the value $y = 10$ at $x_1 = x_2 = 0.5$ is a minimum of $f(x_1, x_2)$.

2.6 Taylor Series

If a function $f(x)$ has infinitely many derivatives (most of which may, however, be zero) near the value $x = x_0$, then the function can be decomposed into the *Taylor series*

$$f(x) = f(x_0) + \frac{f'(x_0)}{1!}(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \frac{f'''(x_0)}{3!}(x - x_0)^3 + \cdots$$
$$= \sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n \tag{2.1}$$

where $f^{(n)}$ represents the nth-order derivative of f, and $n!$ is the *factorial* of n.[8] Taylor series are named after the 18th-century British mathematician Brook Taylor.

As long as x is sufficiently close to x_0, and as long as the function $f(\cdot)$ is sufficiently well behaved, $f(x)$ may be *approximated* by taking only the first few terms of the Taylor series. For example, if the function is nearly quadratic between x and x_0, then $f(x)$ can be approximated by the first three terms of the Taylor expansion, because the remaining derivatives will be small; similarly, if the function is nearly linear between x and x_0, then $f(x)$ can be approximated by the first two terms.

To illustrate the application of Taylor series, consider the cubic function

$$f(x) = 1 + x^2 + x^3$$

Then

$$f'(x) = 2x + 3x^2$$
$$f''(x) = 2 + 6x$$

[8] The factorial of a non-negative integer n is defined as $n! \equiv n(n-1)(n-2)\cdots(2)(1)$; by convention, 0! and 1! are both taken to be 1.

$$f'''(x) = 6$$
$$f^{(n)}(x) = 0 \text{ for } n > 3$$

Let us take $x_0 = 2$; evaluating the function and its derivatives at this value of x,

$$f(2) = 1 + 2^2 + 2^3 = 13$$
$$f'(2) = 2(2) + 3(2)^2 = 16$$
$$f''(2) = 2 + 6(2) = 14$$
$$f'''(2) = 6$$

Finally, let us evaluate $f(x)$ at $x = 4$ using the Taylor-series expansion of the function around $x_0 = 2$:

$$\begin{aligned} f(4) &= f(2) + \frac{f'(2)}{1!}(4-2) + \frac{f''(2)}{2!}(4-2)^2 + \frac{f'''(2)}{3!}(4-2)^3 \\ &= 13 + 16(2) + \frac{14}{2}(2^2) + \frac{6}{6}(2^3) \\ &= 81 \end{aligned}$$

Checking by evaluating the function directly,

$$f(4) = 1 + 4^2 + 4^3 = 1 + 16 + 64 = 81\checkmark$$

In this case, using fewer than all four terms would produce a poor approximation (because, of course, the function in cubic).

Taylor series expansions and approximations generalize to functions of several variables, most simply when the function is scalar-valued and when we can use a first- or second-order approximation. Suppose that $y = f(x_1, x_2, \ldots, x_n) = f(\mathbf{x})$, and that we want to approximate $f(\mathbf{x})$ near the value $\mathbf{x} = \mathbf{x}_0$. Then the secord-order Taylor-series approximation of $f(\mathbf{x})$ is

$$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \left[\mathbf{g}(\mathbf{x}_0)\right]'(\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)'\mathbf{H}(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)$$

where $\mathbf{g}(\mathbf{x}) \equiv \partial y/\partial \mathbf{x}$ and $\mathbf{H}(\mathbf{x}) \equiv \partial^2 y/\partial \mathbf{x}\,\partial \mathbf{x}'$ are, respectively, the gradient and Hessian of $f(\mathbf{x})$, both evaluated at $\mathbf{x}_0$. Notice the strong analogy to the first three terms of the scalar Taylor expansion, given in Equation 2.1.

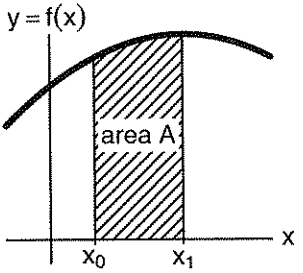

Figure 2.18 The area A under a function $f(x)$ between x_0 and x_1.

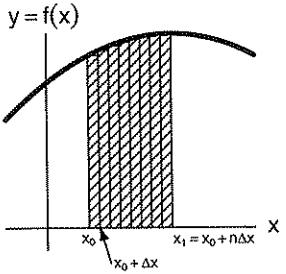

Figure 2.19 Approximating the area under a curve by summing the areas of rectangles.

2.7 Essential Ideas of Integral Calculus

2.7.1 Areas: Definite Integrals

Consider the area A under a curve $f(x)$ between two horizontal coordinates, x_0 and x_1, as illustrated in Figure 2.18. This area can be approximated by dividing the line segment between x_0 and x_1 into n small intervals, each of width Δx, and constructing a series of rectangles just touching the curve, as shown in Figure 2.19. The x-values defining the rectangles are

$$x_0, x_0 + \Delta x, x_0 + 2\Delta x, \ldots, x_0 + n\Delta x$$

Consequently, the combined area of the rectangles is

$$\sum_{i=0}^{n-1} f(x_0 + i\Delta x)\Delta x \approx A$$

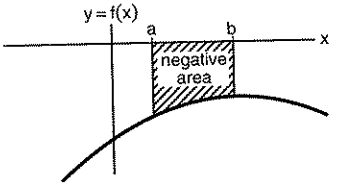

Figure 2.20 The integral $\int_a^b f(x)dx$ is negative because the y values are negative between the limits of integration a and b.

The approximation grows better as the number of rectangles n increases (and Δx grows smaller). In the limit,[9]

$$A = \lim_{\substack{\Delta x \to 0 \\ n \to \infty}} \sum_{i=0}^{n-1} f(x_0 + i\Delta x)\Delta x$$

The following notation is used for this limit, which is called the *definite integral* of $f(x)$ from $x = x_0$ to x_1:

$$A = \int_{x_0}^{x_1} f(x)dx$$

Here, x_0 and x_1 give the *limits of integration*, while the differential dx is the infinitesimal remnant of the width of the rectangles Δx. The symbol for the integral, $\int$, is an elongated "S," indicative of the interpretation of the definite integral as the continuous analog of a sum.

The definite integral defines a *signed area*, which may be negative if (some) values of y are less than 0, as illustrated in Figure 2.20.

2.7.2 Indefinite Integrals

Suppose that for the function $f(x)$, there exists a function $F(x)$ such that

$$\frac{dF(x)}{dx} = f(x)$$

That is, $f(x)$ is the derivative of $F(x)$. Then $F(x)$ is called an *antiderivative* or *indefinite integral* of $f(x)$.

[9]This approach, called *the method of exhaustion* (though not the formal notion of a limit), was known to the ancient Greeks.

The indefinite integral of a function is not unique, for if $F(x)$ is an antiderivative of $f(x)$, then so is $G(x) = F(x) + c$, where c is an arbitrary constant (i.e., not a function of x). Conversely, if $F(x)$ and $G(x)$ are both antiderivatives of $f(x)$, then for some constant c, $G(x) = F(x) + c$.

For example, for $f(x) = x^3$, the function $\frac{1}{4}x^4 + 10$ is an antiderivative of $f(x)$, as are $\frac{1}{4}x^4 - 10$ and $\frac{1}{4}x^4$. Indeed, any function of the form $F(x) = \frac{1}{4}x^4 + c$ will do.

The following notation is used for the indefinite integral: If

$$\frac{dF(x)}{dx} = f(x)$$

then we write

$$F(x) = \int f(x)dx$$

where the integral sign appears without limits of integration. That the same symbol is employed for both areas and antiderivatives (i.e., for definite and indefinite integrals), and that both of these operations are called "integration," are explained in the following section. Notice that while a definite integral—an area—is a particular number, an indefinite integral is a function.

2.7.3 The Fundamental Theorem of Calculus

Newton and Leibniz figured out the connection between antiderivatives and areas under curves. The relationship that they discovered between indefinite and definite integrals is called the *fundamental theorem of calculus*:

$$\int_{x_0}^{x_1} f(x)dx = F(x_1) - F(x_0)$$

where $F(\cdot)$ is *any* antiderivative of $f(\cdot)$.

Here is a nonrigorous proof of this theorem: Consider the area $A(x)$ under the curve $f(x)$ between some fixed value x_0 and another (moveable) value x, as shown in Figure 2.21. The notation $A(x)$ emphasizes that the area is a function of x: As we move x left or right, the area $A(x)$ changes. In Figure 2.21, $x + \Delta x$ is a value slightly to the right of x, and ΔA is the area under the curve between x and $x + \Delta x$. A rectangular approximation to this small area is

$$\Delta A \approx f(x)\Delta x$$

The area ΔA is also

$$\Delta A = A(x + \Delta x) - A(x)$$

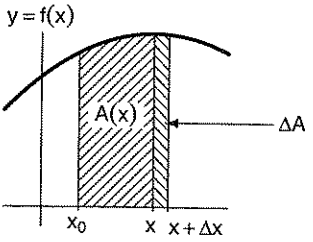

Figure 2.21 The area $A(x)$ under the curve between the fixed value x_0 and another value x.

Taking the derivative of A,

$$\begin{aligned}\frac{dA(x)}{dx} &= \lim_{\Delta x \to 0} \frac{\Delta A}{\Delta x} \\ &= \lim_{\Delta x \to 0} \frac{f(x)\Delta x}{\Delta x} \\ &= f(x)\end{aligned}$$

Consequently,

$$A(x) = \int f(x)dx$$

is a *specific*, but as yet unknown, indefinite integral of $f(x)$. Let $F(x)$ be some *other* specific, arbitrary, indefinite integral of $f(x)$. Then $A(x) = F(x) + c$, for some c (because, as we previously discovered, two indefinite integrals of the same function differ by a constant). We know that $A(x_0) = 0$, because $A(x)$ is the area under the curve between x_0 and any x, and the area under the curve between x_0 and x_0 is 0. Thus,

$$\begin{aligned}A(x_0) &= F(x_0) + c = 0 \\ c &= -F(x_0)\end{aligned}$$

and, for a particular value of $x = x_1$,

$$A(x_1) = \int_{x_0}^{x_1} f(x)dx = F(x_1) - F(x_0)$$

where (recall) $F(\cdot)$ is *an arbitrary* antiderivative of $f(\cdot)$.

For example, let us find the area (evaluate the definite integral)

$$A = \int_1^3 (x^2 + 3)dx$$

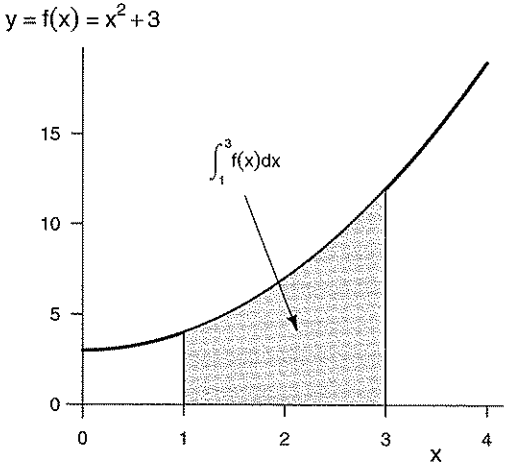

Figure 2.22 The area $A = \int_1^3 (x^2 + 3)dx$.

This area is graphed in Figure 2.22. It is convenient to use[10]

$$F(x) = \frac{1}{3}x^3 + 3x$$

Then

$$\begin{aligned} A &= F(3) - F(1) \\ &= \left(\frac{1}{3}3^3 + 3 \times 3\right) - \left(\frac{1}{3}1^3 + 3 \times 1\right) \\ &= 18 - 3\tfrac{1}{3} = 14\tfrac{2}{3} \end{aligned}$$

2.8 Recommended Reading

There is an almost incredible profusion of introductory calculus texts, and I cannot claim to have read more than a few of them. Of these, my favorite is Thompson and Gardner (1998). For an extensive treatment of calculus of several variables with a social science (specifically, economic) focus, see Binmore and Davies (2001).

[10]Reader: Verify that $F(x)$ is an antiderivative of $f(x) = x^2 + 3$. More generally, one can find antiderivatives of polynomial functions by working the rule for differentiating powers in reverse.

CHAPTER 3. PROBABILITY AND ESTIMATION

The purpose of this chapter is to outline basic results in probability and statistical inference that are employed widely in applied statistics. For good reason, elementary statistics courses—particularly in the social sciences—often provide only the barest introduction to probability and the theory of estimation. To progress past a certain point, however, some background in these topics is necessary.

In Section 3.1, I review concepts in elementary probability theory. Sections 3.2 and 3.3 briefly describe a number of probability distributions that are of special importance in statistics. Section 3.4 outlines asymptotic distribution theory, which is often required to determine properties of statistical estimators, a subject that is taken up in Section 3.5. Section 3.6 develops the broadly applicable and centrally important method of maximum-likelihood estimation. The concluding section of the chapter, Section 3.7, introduces Bayesian estimation. Taken together, the sections of this chapter provide a "crash course" in some of the basics of mathematical statistics.

3.1 Elementary Probability Theory

3.1.1 Probability Basics

In probability theory, an *experiment* is a repeatable procedure for making an observation; an *outcome* is a possible observation resulting from an experiment; and the *sample space* of the experiment is the set of all possible outcomes. Any specific *realization* of the experiment produces a particular outcome in the sample space. Sample spaces may be discrete and finite, discrete and infinite, or continuous.

If, for example, we flip a coin *twice* and record on each flip whether the coin shows heads (H) or tails (T), then the sample space of the experiment is discrete and finite, consisting of the outcomes $S = \{HH, HT, TH, TT\}$. If, alternatively, we flip a coin repeatedly until a head appears, and record the number of flips required to obtain this result, then the sample space is discrete and infinite, consisting of the positive integers, $S = \{1, 2, 3, \ldots\}$.[1]

[1] The sample space is infinite because we may have to wait arbitrarily long to observe the first head, even though a very long wait may be highly improbable. More generally, when S is

If we burn a light bulb until it fails, recording the burning time in hours and fractions of an hour, then the sample space of the experiment is continuous and consists of all positive real numbers (not bothering to specify an upper limit for the life of a bulb): $S = \{x\colon\ x > 0\}$. In this section, I will limit consideration to discrete, finite sample spaces.

An *event* is a subset of the sample space of an experiment—that is, a set of outcomes. An event is said to *occur* in a realization of the experiment if one of its constituent outcomes occurs. For example, for $S = \{HH, HT, TH, TT\}$, the event $E \equiv \{HH, HT\}$, representing a head on the first flip of the coin, occurs if we obtain either the outcome HH or the outcome HT. Note that the sample space S itself, and the *null* or *empty event* $\emptyset \equiv \{\}$, which contains no outcomes, are both events by this definition.

Axioms of Probability Let $S = \{o_1, o_2, \ldots, o_n\}$ be the sample space of an experiment; let $O_1 \equiv \{o_1\}$, $O_2 \equiv \{o_2\}, \ldots, O_n \equiv \{o_n\}$ be the *simple events*, each consisting of one of the outcomes; and let the event $E = \{o_a, o_b, \ldots, o_m\}$ be any subset of S (where subscripts $a, b, \ldots, m$ are different numbers between 1 and n). *Probabilities* are real numbers assigned to events in a manner consistent with the following axioms (rules):[2]

P1: $\Pr(E) \geq 0$: The probability of an event is non-negative.
P2: $\Pr(E) = \Pr(O_a) + \Pr(O_b) + \cdots + \Pr(O_m)$: The probability of an event is the sum of probabilities of its constituent outcomes.
P3: $\Pr(S) = 1$ and $\Pr(\emptyset) = 0$: The sample space is exhaustive—some outcome must occur.

Suppose, for example, that all outcomes in the sample space $S = \{HH, HT, TH, TT\}$ are equally likely, so that

$$\Pr(HH) = \Pr(HT) = \Pr(TH) = \Pr(TT) = .25$$

Then, for $E \equiv \{HH, HT\}$, $\Pr(E) = .25 + .25 = .5$. Equally likely outcomes produce a simple example—and correspond to a "fair" coin "fairly" flipped—but any assignment of probabilities to outcomes that sum to 1 is consistent with the axioms.

---

infinite but discrete, we say that it is *countably infinite*, because a one-to-one correspondence can be established between the elements of S and the natural numbers $0, 1, 2, \ldots$.

[2]These axioms are similar to (and equivalent to) those proposed by the 20th-century Russian mathematician, Andrey Kolmogorov.

In *classical statistics*, the perspective adopted in most applications of statistics, probabilities are interpreted as long-run proportions. Thus, if the probability of an event is $\frac{1}{2}$, then the event will occur approximately half the time when the experiment is repeated many times, and the approximation is expected to improve as the number of repetitions increases. This is sometimes termed an *objective* or *frequentist* interpretation of probability: Probabilities are interpreted as long-run relative frequencies—that is, proportions. (Cf., Section 3.7 on Bayesian statistical inference.)

Relations Among Events, Conditional Probability, and Independence

A number of important relations can be defined among events. The *intersection* of two events, E_1 and E_2, denoted $E_1 \cap E_2$, contains all outcomes common to the two; $\Pr(E_1 \cap E_2)$ is thus the probability that *both* E_1 *and* E_2 occur simultaneously. If $E_1 \cap E_2 = \emptyset$, then E_1 and E_2 are said to be *disjoint* or *mutually exclusive*. By extension, the intersection of many events $E_1 \cap E_2 \cap \cdots \cap E_k$ contains all outcomes that are members of each and every event. Consider, for example, the events $E_1 \equiv \{HH, HT\}$ (a head on the first trial), $E_2 \equiv \{HH, TH\}$ (a head on the second trial), and $E_3 \equiv \{TH, TT\}$ (a tail on the first trial). Then $E_1 \cap E_2 = \{HH\}$, $E_1 \cap E_3 = \emptyset$, and $E_2 \cap E_3 = \{TH\}$.

The *union* of two events $E_1 \cup E_2$ contains all outcomes that are in either or both events; $\Pr(E_1 \cup E_2)$ is the probability that E_1 occurs *or* that E_2 occurs (or that *both* occur). The union of several events $E_1 \cup E_2 \cup \cdots \cup E_k$ contains all outcomes that are in one or more of the events. If these events are disjoint, then

$$\Pr(E_1 \cup E_2 \cup \cdots \cup E_k) = \sum_{i=1}^{k} \Pr(E_i)$$

otherwise

$$\Pr(E_1 \cup E_2 \cup \cdots \cup E_k) < \sum_{i=1}^{k} \Pr(E_i)$$

(because some outcomes contribute more than once when the probabilities are summed). For two events,

$$\Pr(E_1 \cup E_2) = \Pr(E_1) + \Pr(E_2) - \Pr(E_1 \cap E_2)$$

Subtracting the intersection corrects for double counting. To extend the previous example, assuming equally likely outcomes (where, recall, events E_1 and E_3 are disjoint, but E_1 and E_2 are not),

$$\begin{aligned}
\Pr(E_1 \cup E_3) &= \Pr(HH, HT, TH, TT) = 1 \\
&= \Pr(E_1) + \Pr(E_3) \\
&= .5 + .5 \\
\Pr(E_1 \cup E_2) &= \Pr(HH, HT, TH) = .75 \\
&= \Pr(E_1) + \Pr(E_2) - \Pr(E_1 \cap E_2) \\
&= .5 + .5 - .25
\end{aligned}$$

The *conditional probability of* E_2 *given* E_1 is

$$\Pr(E_2|E_1) \equiv \frac{\Pr(E_1 \cap E_2)}{\Pr(E_1)} \tag{3.1}$$

The conditional probability is interpreted as the probability that E_2 will occur if E_1 is known to have occurred. Solving Equation 3.1 for $\Pr(E_1 \cap E_2)$ produces the general *multiplication rule* for probabilities:

$$\Pr(E_1 \cap E_2) = \Pr(E_1) \Pr(E_2|E_1)$$

We can exchange the roles of E_1 and E_2 in these formulas:

$$\Pr(E_1|E_2) \equiv \frac{\Pr(E_1 \cap E_2)}{\Pr(E_2)} \tag{3.2}$$

$$\Pr(E_1 \cap E_2) = \Pr(E_2) \Pr(E_1|E_2) \tag{3.3}$$

Two events are *independent* if $\Pr(E_1 \cap E_2) = \Pr(E_1) \Pr(E_2)$ (the multiplication rule for probabilities of independent events). Independence of E_1 and E_2 implies that $\Pr(E_1) = \Pr(E_1|E_2)$ and that $\Pr(E_2) = \Pr(E_2|E_1)$: That is, the *unconditional probability* of each of two independent events is the same as the conditional probability of that event given the other. More generally, a set of events $\{E_1, E_2, \ldots, E_k\}$ is independent if, for every subset $\{E_a, E_b, \ldots, E_m\}$ containing two or more of the events,

$$\Pr(E_a \cap E_b \cap \cdots \cap E_m) = \Pr(E_a) \Pr(E_b) \cdots \Pr(E_m)$$

Appealing once more to our example, the probability of a head on the second trial (E_2) given a head on the first trial (E_1) is

$$\begin{aligned}
\Pr(E_2|E_1) &= \frac{\Pr(E_1 \cap E_2)}{\Pr(E_1)} \\
&= \frac{.25}{.5} = .5 \\
&= \Pr(E_2)
\end{aligned}$$

Likewise, $\Pr(E_1 \cap E_2) = .25 = \Pr(E_1)\Pr(E_2) = .5 \times .5$. The events E_1 and E_2 are, therefore, independent.

Independence is different from disjointness: If two events are disjoint, then they cannot occur together, and they are, therefore, *dependent*. In our example, the events E_1 and E_2 are independent, not disjoint: $E_1 \cap E_2 = \{HH\} \neq \emptyset$.

The *difference* between two events $E_1 - E_2$ contains all outcomes in the first event that are not in the second. The difference $\overline{E} \equiv S - E$ is called the *complement* of the event E. Note that $\Pr(\overline{E}) = 1 - \Pr(E)$. From the example, where $E_1 \equiv \{HH, HT\}$ with all outcomes equally likely, $\Pr(\overline{E}_1) = \Pr(TH, TT) = .5 = 1 - .5$.

Bonferroni Inequalities Let $E \equiv E_1 \cap E_2 \cap \cdots E_k$. Then $\overline{E} = \overline{E}_1 \cup \overline{E}_2 \cup \cdots \cup \overline{E}_k$. Applying previous results,

$$\begin{aligned} \Pr(E_1 \cap E_2 \cap \cdots \cap E_k) &= \Pr(E) = 1 - \Pr(\overline{E}) \\ &\geq 1 - \sum_{i=1}^{k} \Pr(\overline{E}_i) \end{aligned} \tag{3.4}$$

Suppose that all of the events $E_1, E_2, \ldots, E_k$ have equal probabilities, say $\Pr(E_i) = 1 - b$ [so that $\Pr(\overline{E}_i) = b$]. Then

$$\begin{aligned} \Pr(E_1 \cap E_2 \cap \cdots \cap E_k) &\equiv 1 - a \\ &\geq 1 - kb \end{aligned} \tag{3.5}$$

Equation 3.5 and the more general Equation 3.4 are called *Bonferroni inequalities*, named after Carlo Emilio Bonferroni, a 20th-century Italian mathematician.

Equation 3.5 has the following application to simultaneous statistical inference: Suppose that b is the Type I error rate (i.e., the "α-level") for *each* of k nonindependent statistical tests. Let a represent the combined Type I error rate for the k tests—that is, the probability of falsely rejecting *at least one* of k true null hypotheses. Then $a \leq kb$. For instance, if we test 20 true statistical hypotheses, each at a significance level of .01, then the probability of rejecting *at least one* hypothesis is at most $20 \times .01 = .20$ (i.e., no more than one chance in five)—a sober reminder that naïve "data dredging" can prove seriously misleading.

3.1.2 Random Variables

A *random variable* is a function that assigns a number to each outcome of the sample space of an experiment. For the sample space $S = \{HH, HT, TH, TT\}$, introduced earlier, a random variable X that counts the number of heads in an outcome is defined as follows:

Outcome	*Value x of X*
HH	2
HT	1
TH	1
TT	0

If, as in this example, X is a discrete random variable, then we write $p(x)$ for $\Pr(X = x)$, where the uppercase letter X represents the random variable, while the lowercase letter x denotes a *particular value* of the variable.[3] The probabilities $p(x)$ for all values of X comprise the *probability distribution* (or *probability mass function*) of the random variable. If, for example, each of the four outcomes of the coin-flipping experiment has probability .25, then the probability distribution of the number of heads is

	x	$p(x)$
$\{TT\} \Longrightarrow$	0	.25
$\{HT, TH\} \Longrightarrow$	1	.50
$\{HH\} \Longrightarrow$	2	.25
	Sum	1.00

The table shows the events that map into each value x of the random variable.

The *cumulative distribution function* (*CDF*) of a random variable X, written $P(x)$, gives the probability of observing a value of the variable that is less than or equal to a particular value:

$$P(x) \equiv \Pr(X \leq x) = \sum_{x' \leq x} p(x')$$

[3] A random variable X is discrete if it takes on a finite or countably infinite number of distinct values.

For the example,

x	$P(x)$
0	.25
1	.75
2	1.00

Random variables defined on continuous sample spaces may themselves be continuous. We still take $P(x)$ as $\Pr(X \leq x)$, but it generally becomes meaningless to refer to the probability of observing *individual values* of X.[4] The *probability density function* $p(x)$ is, nevertheless, the continuous analog of the discrete probability distribution, defining $p(x) \equiv dP(x)/dx$.[5] Reversing this relation,[6] $P(x) = \int_{-\infty}^{x} p(x)\,dx$; and

$$\Pr(x_0 \leq X \leq x_1) = P(x_1) - P(x_0) = \int_{x_0}^{x_1} p(x)\,dx$$

Thus, as illustrated in Figure 3.1, areas under the density function represent probabilities.[7]

A particularly simple continuous probability distribution is the *rectangular* (or *uniform*) *distribution*:

$$p(x) = \begin{cases} 0 & a > x \\ \dfrac{1}{b-a} & a \leq x \leq b \\ 0 & x > b \end{cases} \tag{3.6}$$

[4] As explained immediately below, probabilities correspond to *areas* under the density function $p(x)$, and the area above a *particular* value of x_0 of X (i.e., a vertical line) is 0.

[5] The probability density function of a continuous random variable X is more commonly denoted $f(x)$, and its cumulative distribution function $F(x)$, but I find $p(x)$ and $P(x)$ more natural, and prefer to reserve $f(\cdot)$ for other purposes, such as transformations of random variables (see Section 3.1.3).

[6] If you are unfamiliar with integral calculus (which is described in Section 2.7), do not be too concerned: The principal point to understand is that *areas* under the density curve $p(x)$ are interpreted as probabilities, and that the *height* of the CDF $P(x)$ gives the probability of observing values of X less than or equal to the value x. The integral sign $\int$ is the continuous analog of a sum, and represents the area under a curve.

[7] Because of the continuity of $p(x)$, and the associated fact that $\Pr(X = x_0) = \Pr(X = x_1) = 0$, we need not distinguish between $\Pr(x_0 \leq X \leq x_1)$ and $\Pr(x_0 < X < x_1)$.

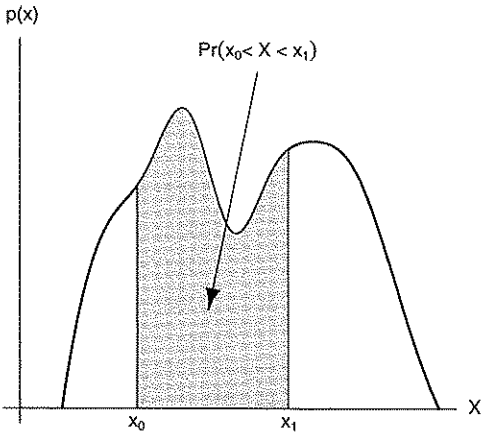

Figure 3.1 Areas under the probability density function $p(x)$ are probabilities.

This density function is pictured in Figure 3.2(a), and the corresponding cumulative distribution function is shown in Figure 3.2(b). The total area under a density function must be 1; here,

$$\int_{-\infty}^{\infty} p(x)dx = \int_{a}^{b} p(x)dx = \frac{1}{b-a}(b-a) = 1$$

The *support* of a random variable is the set of values for which the probability or probability density is nonzero; the support of the rectangular distribution is therefore $a \leq X \leq b$.

Two fundamental properties of a random variable are its *expected value* (or *mean*) and its *variance*.[8] The expected value specifies the center of the probability distribution of the random variable (in the same sense as the mean of a set of scores specifies the center of their distribution), while the variance indicates how spread out the distribution is around its expectation. The expectation is interpretable as the mean score of the random variable that would be observed over many repetitions of the experiment, while the variance is the mean-squared distance between the scores and their expectation.

[8]The expectation and variance are undefined for some random variables, a possibility that I will ignore here.

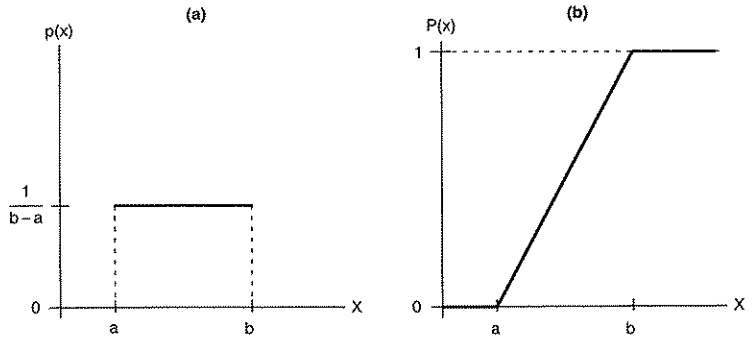

Figure 3.2 (a) The probability density function $p(x)$, and (b) the cumulative distribution function $P(x)$ for the rectangular distribution.

In the discrete case, the expectation of a random variable X, symbolized by $E(X)$ or μ_X, is given by

$$E(X) \equiv \sum_{\text{all } x} xp(x)$$

The analogous formula for the continuous case is[9]

$$E(X) \equiv \int_{-\infty}^{\infty} xp(x)\,dx$$

The variance of a random variable X, written $V(X)$ or σ_X^2, is defined as

$$\begin{aligned} V(X) &\equiv E[(X - \mu_X)^2] \\ &= E(X^2) - \mu_X^2 \end{aligned}$$

Thus, in the discrete case,

$$V(X) \equiv \sum_{\text{all } x} (x - \mu_X)^2 p(x)$$

while, in the continuous case,

$$V(X) \equiv \int_{-\infty}^{\infty} (x - \mu_X)^2 p(x)\,dx$$

[9] We need only integrate over the support of X, which may not include the entire real line.

The variance is expressed in the squared units of the random variable (e.g., "squared number of heads"), but the *standard deviation* $\sigma \equiv +\sqrt{\sigma^2}$ is measured in the same units as the variable.

For our example,

x	$p(x)$	$xp(x)$	$x-\mu$	$(x-\mu)^2 p(x)$
0	.25	0.00	−1	0.25
1	.50	0.50	0	0.00
2	.25	0.50	1	0.25
Sum	1.00	$\mu = 1.00$		$\sigma^2 = 0.50$

Thus, $E(X) = 1$, $V(X) = 0.5$, and $\sigma = \sqrt{0.5} \approx 0.707$. Similarly, for the rectangular distribution (Equation 3.6),

$$E(X) = \int_a^b x\left(\frac{1}{b-a}\right)dx = \frac{a+b}{2}$$

$$V(X) = \int_a^b \left(x - \frac{a+b}{2}\right)^2 \left(\frac{1}{b-a}\right)dx = \frac{(a-b)^2}{12}$$

The *joint probability distribution* of two discrete random variables X_1 and X_2 gives the probability of simultaneously observing any pair of values for the two variables. We write $p_{12}(x_1, x_2)$ for $\Pr(X_1 = x_1 \text{ and } X_2 = x_2)$; it is usually unambiguous to drop the subscript on p, simply writing $p(x_1, x_2)$. The *joint probability density* $p(x_1, x_2)$ of two continuous random variables is defined analogously. Extension to the joint probability or joint probability density $p(x_1, x_2, \ldots, x_n)$ of several random variables is straightforward.

To distinguish it from the joint probability distribution, we call $p_1(x_1)$ the *marginal probability distribution* or *marginal probability density* for X_1. Note that $p_1(x_1) = \sum_{x_2} p(x_1, x_2)$ or $p_1(x_1) = \int_{-\infty}^{\infty} p(x_1, x_2)\, dx_2$. We usually drop the subscript, to write $p(x_1)$.

In the fair coin-flipping experiment, for example, let X_1 count the number of heads, and let $X_2 = 1$ if both coins are the same and 0 if they are different:

Outcome	*Pr*	x_1	x_2
HH	.25	2	1
HT	.25	1	0
TH	.25	1	0
TT	.25	0	1

The joint and marginal distributions for X_1 and X_2 are as follows:

$p(x_1, x_2)$			
	x_2		
x_1	0	1	$p(x_1)$
0	0	.25	.25
1	.50	0	.50
2	0	.25	.25
$p(x_2)$	.50	.50	1.00

The *conditional probability* or *conditional probability density* of X_1 given X_2 is

$$p_{1|2}(x_1|x_2) = \frac{p_{12}(x_1, x_2)}{p_2(x_2)}$$

As before, it is generally convenient to drop the subscript, writing $p(x_1|x_2)$. For our example, the conditional distributions $p(x_1|x_2)$ for $x_2 = 0$ and $x_2 = 1$ are

$p(x_1\|x_2)$		
	x_2	
x_1	0	1
0	0	.5
1	1.0	0
2	0	.5
Sum	1.0	1.0

The *conditional expectation* of X_1 given $X_2 = x_2$—written $E_{1|2}(X_1|x_2)$ or, more compactly, $E(X_1|x_2)$—is found from the conditional distribution $p_{1|2}(x_1|x_2)$, as is the *conditional variance* of X_1 given $X_2 = x_2$, written $V_{1|2}(X_1|x_2)$ or $V(X_1|x_2)$; for example, in the discrete case,

$$E_{1|2}(X_1|x_2) = \sum_{x_1} x_1 p_{1|2}(x_1|x_2)$$

$$V_{1|2}(X_1|x_2) = \sum_{x_1} \left[x_1 - E_{1|2}(X_1|x_2)\right]^2 p_{1|2}(x_1|x_2)$$

Using the illustrative conditional distributions $p_{12}(x_1|x_2)$,

$$E_{1|2}(X_1|0) = 0(0) + 1(1) + 0(2) = 1$$

$$V_{1|2}(X_1|0) = 0(0-1)^2 + 1(1-1)^2 + 0(2-1)^2 = 0$$

$$E_{1|2}(X_1|1) = .5(0) + 0(1) + .5(2) = 1$$
$$V_{1|2}(X_1|1) = .5(0-1)^2 + 0(1-1)^2 + .5(2-1)^2 = 1$$

The random variables X_1 and X_2 are said to be *independent* if $p(x_1) = p(x_1|x_2)$ for all values of X_1 and X_2; that is, when X_1 and X_2 are independent, the marginal and conditional distributions of X_1 are identical. Equivalent conditions for independence are $p(x_2) = p(x_2|x_1)$ and $p(x_1, x_2) = p(x_1)p(x_2)$: When X_1 and X_2 are independent, their joint probability or probability density is the product of their marginal probabilities or densities. In our example, it is clear that X_1 and X_2 are *not* independent. More generally, the set of n random variables $\{X_1, X_2, \ldots, X_n\}$ is independent if for every subset $\{X_a, X_b, \ldots, X_m\}$ of size $m = 2$ or larger,

$$p(x_a, x_b, \ldots, x_m) = p(x_a)p(x_b)\cdots p(x_m)$$

The *covariance* of two random variables is a measure of their *linear* dependence:

$$C(X_1, X_2) = \sigma_{12} \equiv E[(X_1 - \mu_1)(X_2 - \mu_2)]$$
$$= E(X_1X_2) - \mu_1\mu_2$$

When large values of X_1 are associated with large values of X_2 (and, conversely, small values with small values), the covariance is positive; when large values of X_1 are associated with small values of X_2 (and vice versa), the covariance is negative. The covariance is 0 otherwise, for instance—but not exclusively—when the random variables are independent. That is, two random variables may be *nonlinearly* related and still have covariance 0: In our previous example, X_1 and X_2 are not independent, but σ_{12} is nevertheless 0 (as the reader can verify). The covariance of a variable with itself is its variance: $C(X, X) = V(X)$.

The *correlation* $\rho_{12} \equiv \sigma_{12}/\sigma_1\sigma_2$ between two random variables X_1 and X_2 is a normalized version of the covariance. The smallest possible value of the correlation, $\rho = -1$, is indicative of a perfect inverse linear relationship between the random variables, while the largest value, $\rho = 1$, is indicative of a perfect direct linear relationship; $\rho = 0$ corresponds to a covariance of 0 and indicates the absence of a linear relationship.

Vector Random Variables It is often convenient to write a collection of random variables as a *vector random variable*: for example,

$\underset{(n\times 1)}{\mathbf{x}} = [X_1, X_2, \ldots, X_n]'$. The expectation of a vector random variable is simply the vector of expectations of its elements:

$$E(\mathbf{x}) = \boldsymbol{\mu}_{\mathbf{x}} \equiv [E(X_1), E(X_2), \ldots, E(X_n)]'$$

The *variance-covariance matrix* of a vector random variable $\mathbf{x}$ is defined in analogy to the scalar variance as

$$V(\mathbf{x}) = \underset{(n\times n)}{\boldsymbol{\Sigma}_{\mathbf{xx}}} \equiv E[(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})'] = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{bmatrix}$$

The diagonal entries of $V(\mathbf{x})$ are the variances of the Xs, and the off-diagonal entries are their covariances. The variance-covariance matrix $V(\mathbf{x})$ is symmetric and positive semi-definite (see Section 1.6). The *covariance matrix* of two vector random variables $\underset{(n\times 1)}{\mathbf{x}}$ and $\underset{(m\times 1)}{\mathbf{y}}$ is

$$C(\mathbf{x}, \mathbf{y}) = \underset{(n\times m)}{\boldsymbol{\Sigma}_{\mathbf{xy}}} \equiv E[(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}})'] = \begin{bmatrix} \sigma_{x_1y_1} & \sigma_{x_1y_2} & \cdots & \sigma_{x_1y_m} \\ \sigma_{x_2y_1} & \sigma_{x_2y_2} & \cdots & \sigma_{x_2y_m} \\ \vdots & \vdots & & \vdots \\ \sigma_{x_ny_1} & \sigma_{x_ny_2} & \cdots & \sigma_{x_ny_m} \end{bmatrix}$$

and consists of the covariances between all pairs of Xs and Ys.

3.1.3 Transformations of Random Variables

Suppose that the random variable Y is a linear function $a + bX$ (where a and b are constants) of a discrete random variable X, which has expectation μ_X and variance σ_X^2. Then

$$\begin{aligned} E(Y) = \mu_Y &= \sum_x (a + bx)\, p(x) \\ &= a \sum p(x) + b \sum x p(x) \\ &= a + b\mu_X \end{aligned}$$

and (employing this property of the expectation operator)

$$\begin{aligned} V(Y) &= E[(Y - \mu_Y)^2] = E\{[(a + bX) - (a + b\mu_X)]^2\} \\ &= b^2 E[(X - \mu_X)^2] = b^2\sigma_X^2 \end{aligned}$$

Now, let Y be a linear function $a_1X_1 + a_2X_2$ of two discrete random variables X_1 and X_2, with expectations μ_1 and μ_2, variances σ_1^2 and σ_2^2, and covariance σ_{12}. Then

$$\begin{aligned} E(Y) = \mu_Y &= \sum_{x_1}\sum_{x_2}(a_1x_1 + a_2x_2)p(x_1, x_2) \\ &= \sum_{x_1}\sum_{x_2}a_1x_1p(x_1, x_2) + \sum_{x_1}\sum_{x_2}a_2x_2p(x_1, x_2) \\ &= a_1\sum_{x_1}x_1p(x_1) + a_2\sum_{x_2}x_2p(x_2) \\ &= a_1\mu_1 + a_2\mu_2 \end{aligned}$$

and

$$\begin{aligned} V(Y) &= E[(Y - \mu_Y)^2] \\ &= E\{[(a_1X_1 + a_2X_2) - (a_1\mu_1 + a_2\mu_2)]^2\} \\ &= a_1^2E[(X_1 - \mu_1)^2] + a_2^2E[(X_2 - \mu_2)^2] \\ &\quad + 2a_1a_2E[(X_1 - \mu_1)(X_2 - \mu_2)] \\ &= a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + 2a_1a_2\sigma_{12} \end{aligned}$$

When X_1 and X_2 are independent and, consequently, $\sigma_{12} = 0$, this expression simplifies to $V(Y) = a_1^2\sigma_1^2 + a_2^2\sigma_2^2$.

Although I have developed these rules for discrete random variables, they apply equally to the continuous case. For instance, if $Y = a + bX$ is a linear function of the continuous random variable X, then[10]

$$\begin{aligned} E(Y) &= \int_{-\infty}^{\infty}(a + bx)p(x)\,dx \\ &= a\int_{-\infty}^{\infty}p(x)\,dx + b\int_{-\infty}^{\infty}xp(x)\,dx \\ &= a + bE(X) \end{aligned}$$

Transformations of Vector Random Variables These results generalize to vector random variables in the following manner: Let $\underset{(m\times 1)}{\mathbf{y}}$ be a linear transformation $\underset{(m\times n)(n\times 1)}{\mathbf{A}}\,\mathbf{x}$ of the vector random variable $\mathbf{x}$, which has

[10] If you are unfamiliar with calculus (which is introduced in Chapter 2), then simply think of the integral $\int$ as the continuous analog of the sum $\sum$.

expectation $E(\mathbf{x}) = \boldsymbol{\mu}_\mathbf{x}$ and variance-covariance matrix $V(\mathbf{x}) = \boldsymbol{\Sigma}_{\mathbf{xx}}$. Then it can be shown (in a manner analogous to the scalar proofs given previously) that

$$E(\mathbf{y}) = \boldsymbol{\mu}_\mathbf{y} = \mathbf{A}\boldsymbol{\mu}_\mathbf{x}$$
$$V(\mathbf{y}) = \boldsymbol{\Sigma}_{\mathbf{yy}} = \mathbf{A}\boldsymbol{\Sigma}_{\mathbf{xx}}\mathbf{A}'$$

If the entries of $\mathbf{x}$ are pair-wise independent, then all of the off-diagonal entries of $\boldsymbol{\Sigma}_{\mathbf{xx}}$ are 0, and the variance of each entry of $\mathbf{y}$ takes an especially simple form:

$$\sigma^2_{Y_i} = \sum_{j=1}^{n} a_{ij}^2 \sigma^2_{X_j}$$

At times, when $\mathbf{y} = f(\mathbf{x})$, we need to know not only $E(\mathbf{y})$ and $V(\mathbf{y})$, but also the probability distribution of $\mathbf{y}$. Indeed, the transformation $f(\cdot)$ may be nonlinear. Suppose that there is the same number of elements n in $\mathbf{y}$ and $\mathbf{x}$; that the function f is differentiable; and that f is one to one over the domain of $\mathbf{x}$-values under consideration (i.e., there is a unique pairing of $\mathbf{x}$-values and $\mathbf{y}$-values). This last property implies that we can write the reverse transformation $\mathbf{x} = f^{-1}(\mathbf{y})$. The probability density for $\mathbf{y}$ is given by

$$p(\mathbf{y}) = p(\mathbf{x}) \left| \det \left(\frac{\partial \mathbf{x}}{\partial \mathbf{y}} \right) \right| = p(\mathbf{x}) \left| \det \left(\frac{\partial \mathbf{y}}{\partial \mathbf{x}} \right) \right|^{-1}$$

where $|\det(\partial \mathbf{x}/\partial \mathbf{y})|$, called the *Jacobian* of the transformation, is the absolute value of the $(n \times n)$ determinant

$$\det \begin{bmatrix} \frac{\partial X_1}{\partial Y_1} & \cdots & \frac{\partial X_n}{\partial Y_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial X_1}{\partial Y_n} & \cdots & \frac{\partial X_n}{\partial Y_n} \end{bmatrix}$$

and $|\det(\partial \mathbf{y}/\partial \mathbf{x})|$ is similarly defined. The Jacobian is named after the 19th-century German mathematician Carl Gustav Jacob Jacobi.

3.2 Some Discrete Probability Distributions

In this section, I define several important families of discrete probability distributions: the binomial and Bernoulli distributions; the multinomial

distributions; the Poisson distributions, which can be construed as an approximation to the binomial; and the negative binomial distributions. It is sometimes convenient to refer to a family of probability distributions in the singular—for example, the "binomial *distribution*," rather than the "binomial *distributions*." The discrete probability distributions in this section and the continuous distributions in the next play important roles in statistical inference and statistical modeling.

3.2.1 The Binomial and Bernoulli Distributions

The coin-flipping experiment described at the beginning of Section 3.1.2 gives rise to a binomial random variable that counts the number of heads in two independent flips of a fair coin. To extend this example, let the random variable X count the number of heads in n independent flips of a coin. Let π denote the probability (not necessarily .5) of obtaining a head on any given flip; then $1 - \pi$ is the probability of obtaining a tail.[11] The probability of observing exactly x heads and $n - x$ tails [i.e., $\Pr(X = x)$] is given by the *binomial distribution*

$$p(x) = \binom{n}{x}\pi^x(1-\pi)^{n-x} \tag{3.7}$$

where x is any integer between 0 and n, inclusive; the factor $\pi^x(1-\pi)^{n-x}$ is the probability of observing x heads and $n - x$ tails in a *particular* arrangement; and $\binom{n}{x} \equiv n!/[x!(n-x)!]$, called the *binomial coefficient*, is the number of *different* arrangements of x heads and $n - x$ tails.[12]

The expectation of the binomial random variable X is $E(X) = n\pi$, and its variance is $V(X) = n\pi(1-\pi)$. Figure 3.3 shows the binomial distribution for $n = 10$ and $\pi = .7$. If the products $n\pi$ and $n(1-\pi)$ are sufficiently large (say, both at least equal to 10), then the discrete binomial distribution can be accurately approximated by the continuous normal distribution with the same mean and standard deviation. (The normal distributions are introduced in Section 3.3.1.)

---

[11] The Greek letter π is used because the probability cannot be directly observed. Because π is a probability, it is a number between 0 and 1—*not to be confused with* the mathematical constant ≈ 3.1416.

[12] Recall that the exclamation point is the *factorial* operator:

$$n! \equiv n \times (n-1) \times \cdots \times 2 \times 1 \text{ for integer } n > 1$$
$$\equiv 1 \text{ for } n = 0 \text{ or } 1$$

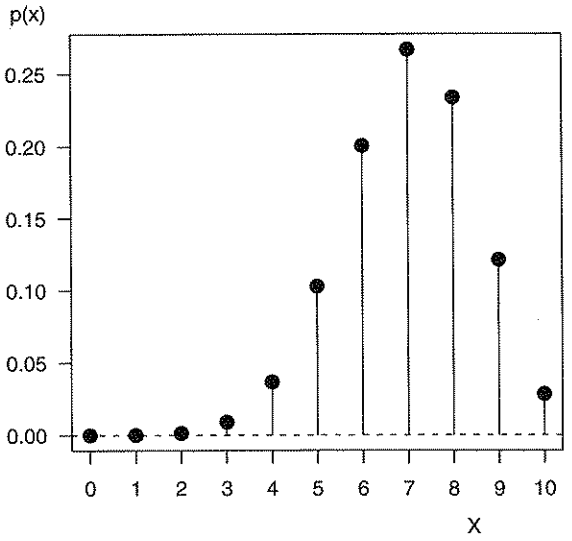

Figure 3.3 The binomial distribution for $n = 10$ and $\pi = .7$.

A binomial random variable is to be distinguished from a *Bernoulli random variable*, which takes on the values 0 and 1 with probabilities π and $1 - \pi$, respectively. The mean and variance of a Bernoulli random variable X are $E(X) = \pi$ and $V(X) = \pi(1 - \pi)$. A Bernoulli random variable could be used to model a *single* flip of a coin, for example, assigning $X = 1$ to a head and $X = 0$ to a tail. The *sum* of independent, identically distributed Bernoulli random variables is therefore binomially distributed. The Bernoulli distribution is named after the 17th-century Swiss mathematician Jacob Bernoulli.

3.2.2 The Multinomial Distributions

Imagine n repeated, independent trials of a process that on each trial can give rise to one of k different categories of outcomes. Let the random variable X_i count the number of outcomes in category i. Let π_i denote the probability of obtaining an outcome in category i on any given trial. Then $\sum_{i=1}^{k} \pi_i = 1$ and $\sum_{i=1}^{k} X_i = n$.

Suppose, for instance, that we toss a die n times, letting X_1 count the number of 1s, X_2 the number of 2s, ..., X_6 the number of 6s. Then $k = 6$,

and π_1 is the probability of obtaining a 1 on any toss, π_2 is the probability of obtaining a 2, and so on. If the die is "fair," then $\pi_1 = \pi_2 = \cdots = \pi_6 = 1/6$.

Returning to the general case, the vector random variable $\mathbf{x} \equiv [X_1, X_2, \ldots, X_k]'$ follows the *multinomial distribution*,

$$p(\mathbf{x}) = p(x_1, x_2, \ldots, x_k) = \frac{n!}{x_1!x_2!\cdots x_k!}\pi_1^{x_1}\pi_2^{x_2}\cdots\pi_k^{x_k}$$

The rationale for this formula is similar to that of the binomial: $\pi_1^{x_1}\pi_2^{x_2}\cdots\pi_k^{x_k}$ gives the probability of obtaining x_1 outcomes in category 1, x_2 in category 2, and so on, in a *particular* arrangement; and $n!/(x_1!x_2!\cdots x_k!)$ counts the number of *different* arrangements. If $k = 2$, then $x_2 = n - x_1$, and the multinomial distribution reduces to the binomial distribution of Equation 3.7 (on page 99).

The expectations of the elements of $\mathbf{x}$ are $E(X_i) = n\pi_i$; their variances are $V(X_i) = n\pi_i(1 - \pi_i)$; and their covariances are $C(X_i, X_j) = -n\pi_i\pi_j$.

3.2.3 The Poisson Distributions

The 19th-century French mathematician Siméon-Denis Poisson introduced the distribution that bears his name as an approximation to the binomial. The approximation is accurate when n is large and π is small, and when the product of the two, $\lambda \equiv n\pi$, is neither large nor small. The Poisson distribution is

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!} \quad \text{for } x = 0, 1, 2, 3, \ldots \text{ and } \lambda > 0$$

Although the domain of X is all non-negative integers, the approximation works because $p(x) \approx 0$ when x is sufficiently large. (Here, $e \approx 2.718$ is the mathematical constant.)

The Poisson distribution arises naturally in several other contexts. Suppose, for example, that we observe a process that randomly produces events of a particular kind (such as births or auto accidents), counting the number of events X that occur in a fixed time interval. This count follows a Poisson distribution if the following conditions hold:

- Although the particular time points at which the events occur are random, the *rate* of occurrence is fixed during the interval of observation

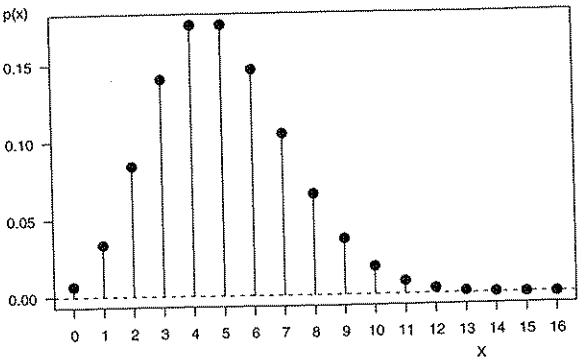

Figure 3.4 The Poisson distribution with rate parameter $\lambda = 5$.

- If we focus attention on a sufficiently small subinterval of length s, then the probability of observing one event in that subinterval is proportional to its length, λs, and the probability of observing more than one event is negligible. In this context, it is natural to think of the parameter λ of the Poisson distribution as the *rate of occurrence* of the event.
- The occurrence of events in nonoverlapping subintervals is independent.

The expectation of a Poisson random variable is $E(X) = \lambda$, and its variance is also $V(X) = \lambda$. Figure 3.4 illustrates the Poisson distribution with rate parameter $\lambda = 5$ (implying that, on average, five events occur during the fixed period of observation).

3.2.4 The Negative Binomial Distributions

Imagine an experiment in which a coin is flipped independently until a fixed "target" number of s *heads* is achieved, and let the random variable X count the number of *tails* that are observed before the target is reached. Then X follows a *negative binomial distribution*, with probability distribution

$$p(x) = \binom{s+x-1}{x} \pi^s (1-\pi)^x \text{ for } x = 0, 1, 2, \ldots$$

where π is the probability of a head on an individual flip of the coin. The expectation of the negative binomial random variable is $E(X) = s(1-\pi)/\pi$,

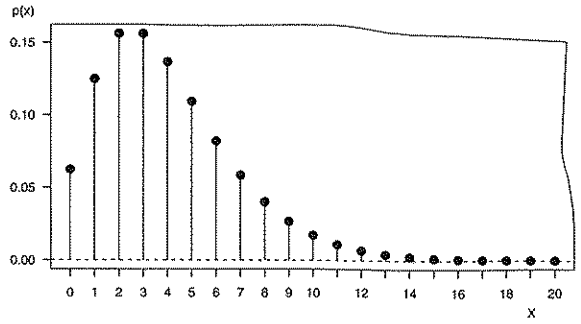

Figure 3.5 Negative binomial distribution for $s = 4$ and $\pi = .5$.

and its variance is $V(X) = s(1 - \pi)/\pi^2$. Figure 3.5 shows the negative binomial distribution for $s = 4$ and $\pi = .5$.

3.3 Some Continuous Distributions

In this section, I describe several important families of continuous distributions: the normal, chi-square, t-, F-, multivariate-normal, exponential, inverse Gaussian, gamma, and beta distributions.

3.3.1 The Normal Distributions

A *normally distributed* (or *Gaussian*) random variable X has probability density function

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \text{ for } -\infty < x < \infty$$

where the parameters of the distribution μ and σ^2 are, respectively, the mean and variance of X. There is, therefore, a different normal distribution for each choice of μ and σ^2; several examples are shown in Figure 3.6. The common abbreviated notation $X \sim N(\mu, \sigma^2)$ means that X is normally distributed with expectation μ and variance σ^2.[13] The Gaussian distributions are named after the great German mathematician Carl Friedrich Gauss

[13]Some authors use the alternative notation $N(\mu, \sigma)$, giving the mean and *standard deviation* of the normally distributed variable rather than its mean and *variance*.

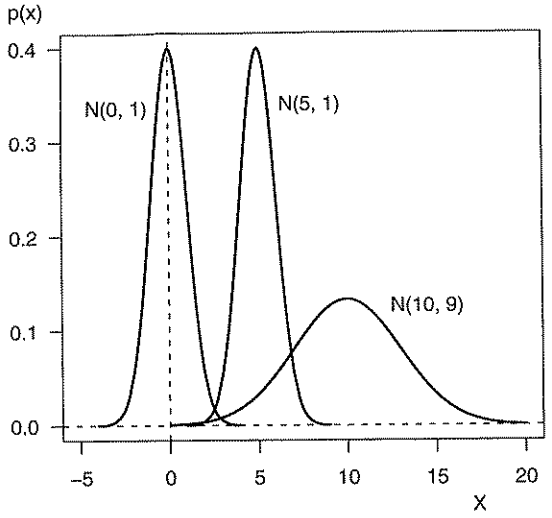

Figure 3.6 Normal density functions: $N(0, 1)$, $N(5, 1)$, and $N(10, 9)$.

(1777–1855), although they were first introduced in 1734 by the French mathematician Abraham de Moivre as an approximation to the binomial distribution.

Of particular importance is the *unit-normal* (or *standard-normal*) random variable $Z \sim N(0, 1)$, with density function

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp(-z^2/2) \text{ for } -\infty < z < \infty$$

The CDF of the unit-normal distribution, $\Phi(z)$, is shown in Figure 3.7. Any normally distributed random variable $X \sim N(\mu, \sigma^2)$ can be transformed to the unit-normal distribution by *standardization*:[14]

$$Z \equiv \frac{X - \mu}{\sigma}$$

[14] *Any* random variable with finite mean and variance can be standardized to mean 0 and variance 1, but standardization leaves the *shape* of the distribution unchanged—in particular, it does not magically transform a non-normal variable to normality.

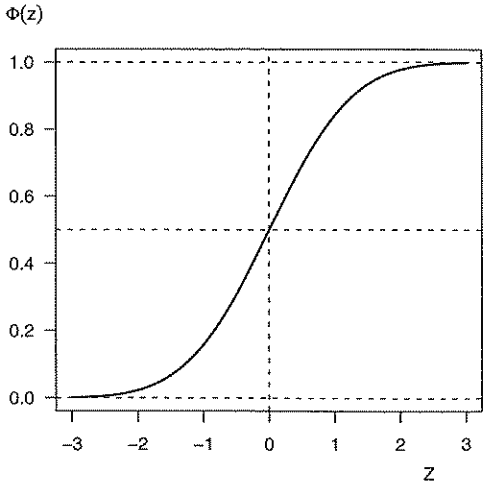

Figure 3.7 The CDF of the unit-normal distribution, $\Phi(z)$.

3.3.2 The Chi-Square (χ^2) Distributions

If $Z_1, Z_2, \ldots, Z_n$ are independently distributed unit-normal random variables, then

$$X^2 \equiv Z_1^2 + Z_2^2 + \cdots + Z_n^2$$

follows a *chi-square distribution* with n *degrees of freedom*, abbreviated χ_n^2. The probability density function of the chi-square variable is

$$P(x^2) = \frac{1}{2^{n/2}\Gamma(\frac{n}{2})}(x^2)^{(n-2)/2}\exp(-x^2/2) \text{ for } x^2 \geq 0$$

where $\Gamma(\cdot)$ is the *gamma function*

$$\Gamma(v) \equiv \int_0^\infty e^{-v} z^{v-1} dz \tag{3.8}$$

(for the generic argument $v > 0$), which is a kind of continuous generalization of the factorial function; in particular, when v is a non-negative integer, $v! = \Gamma(v+1)$. In the current case,

$$\Gamma\left(\frac{n}{2}\right) = \begin{cases} \left(\frac{n}{2}-1\right)! & \text{for } n \text{ even} \\ \left(\frac{n}{2}-1\right)\left(\frac{n}{2}-2\right)\cdots\left(\frac{3}{2}\right)\left(\frac{1}{2}\right)\sqrt{\pi} & \text{for } n \text{ odd} \end{cases}$$

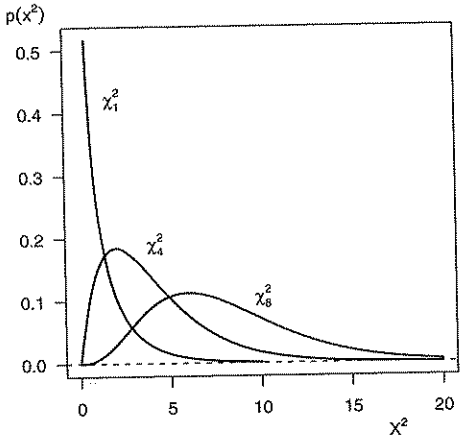

Figure 3.8 Chi-square density functions: χ^2_1, χ^2_4, and χ^2_8.

The expectation and variance of a chi-square random variable are $E(X^2) = n$, and $V(X^2) = 2n$. Several chi-square distributions are graphed in Figure 3.8. As is suggested by the graph, the chi-square distributions are positively skewed, but grow more symmetric (and approach normality) as degrees of freedom increase.

If $X_1^2, X_2^2, \ldots, X_k^2$ are independent chi-square random variables with n_1, $n_2, \ldots, n_k$ degrees of freedom, consecutively, then $X \equiv X_1^2 + X_2^2 + \cdots + X_n^2$ is chi-square distributed with $n = n_1 + n_2 + \cdots + n_k$ degrees of freedom.

3.3.3 Student's t-Distributions

If Z follows a unit-normal distribution, and X^2 independently follows a chi-square distribution with n degrees of freedom, then

$$t \equiv \frac{Z}{\sqrt{\frac{X^2}{n}}}$$

is a *Student's t* random variable with n degrees of freedom, abbreviated t_n.[15] The probability density function of t is

$$p(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\,\Gamma\left(\frac{n}{2}\right)} \times \frac{1}{\left(1 + \frac{t^2}{n}\right)^{(n+1)/2}} \text{ for } -\infty < t < \infty \qquad (3.9)$$

[15] I write a lowercase t for the random variable in deference to nearly universal usage.

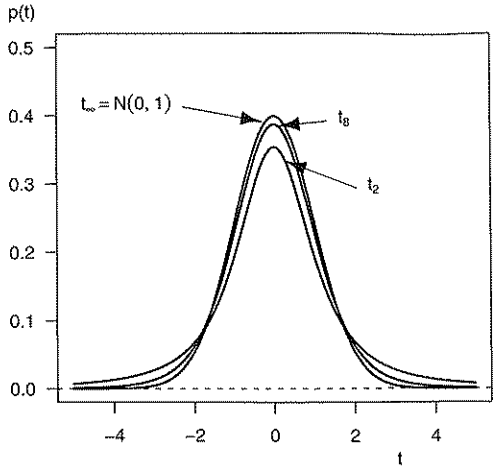

Figure 3.9 t density functions: t_2, t_8, and $t_\infty = N(0, 1)$.

From the symmetry of this formula around $t = 0$, it is clear that $E(t) = 0$.[16] It can be shown that $V(t) = n/(n-2)$, for $n > 2$; thus, the variance of t is large for small degrees of freedom, and approaches 1 as n increases.

Several t-distributions are shown in Figure 3.9. As degrees of freedom grow, the t-distribution approaches the unit-normal distribution, and in the limit, $t_\infty = N(0, 1)$. The normal approximation to the t-distribution is quite close for n as small as 30.

Student's t-distribution is named after William Sealy Gossett, an early 20th-century English statistician employed by the Guiness brewery in Dublin, who wrote under the pen name "Student." Student's t-distribution played an important role in the development of small-sample statistical inference.

3.3.4 The F-Distributions

Let X_1^2 and X_2^2 be independently distributed chi-square variables with n_1 and n_2 degrees of freedom, respectively. Then

$$F \equiv \frac{X_1^2/n_1}{X_2^2/n_2}$$

[16] When $n = 1$, the expectation $E(t)$ does not exist, but the median and mode of t are still 0; t_1 is called the *Cauchy distribution*, named after the 19th-century French mathematician Augustin Louis Cauchy.

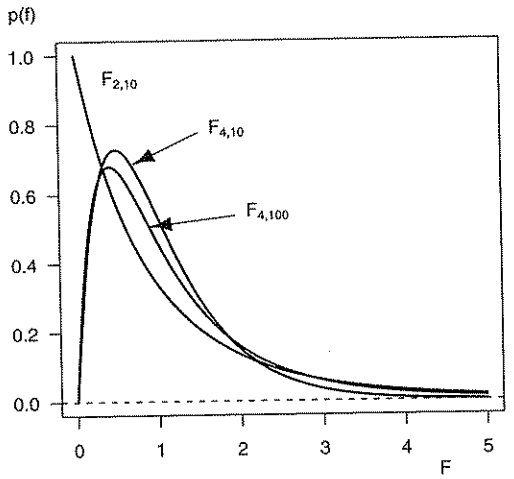

Figure 3.10 F density functions: $F_{2,10}$, $F_{4,10}$, and $F_{4,100}$.

follows an *F-distribution* with n_1 numerator degrees of freedom and n_2 denominator degrees of freedom, abbreviated F_{n_1,n_2}. The F-distribution was named in honor of its discoverer, the great British statistician Sir R. A. Fisher, by the 20th-century American statistician George W. Snedecor.

The probability density for F is

$$p(f)=\frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)}\left(\frac{n_1}{n_2}\right)^{n_1/2} f^{(n_1-2)/2}\left(1+\frac{n_1}{n_2}f\right)^{-(n_1+n_2)/2} \quad \text{for } f \geq 0 \tag{3.10}$$

Comparing Equations 3.9 and 3.10, it can be shown that $t_n^2 = F_{1,n}$. Moreover, as n_2 grows larger, F_{n_1,n_2} approaches $\chi^2_{n_1}/n_1$ and, in the limit, $F_{n,\infty} = \chi^2_n/n$.

For $n_2 > 2$, the expectation of F is $E(F) = n_2/(n_2 - 2)$, which is approximately 1 for large values of n_2. For $n_2 > 4$,

$$V(F) = \frac{2n_2^2(n_1 + n_2 - 2)}{n_1(n_2 - 2)^2(n_2 - 4)}$$

Figure 3.10 shows several F probability density functions. The F-distributions are positively skewed.

3.3.5 The Multivariate-Normal Distributions

The joint probability density for a *multivariate-normal* vector random variable $\mathbf{x} = [X_1, X_2, \ldots, X_n]'$ with mean vector $\boldsymbol{\mu}$ and positive-definite variance-covariance matrix $\boldsymbol{\Sigma}$ is given by

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}\sqrt{\det \boldsymbol{\Sigma}}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right]$$

which is abbreviated as $\mathbf{x} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

If $\mathbf{x}$ is multivariately normally distributed, then the marginal distribution of each of its components is univariate normal, $X_i \sim N(\mu_i, \sigma_i^2)$;[17] and the conditional distribution of any subset of variables given the others, $p(\mathbf{x}_1|\mathbf{x}_2)$, where $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2\}$, is also normal. Furthermore, if $\mathbf{x} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and

$$\underset{(m\times 1)}{\mathbf{y}} = \underset{(m\times n)}{\mathbf{A}}\,\underset{(n\times 1)}{\mathbf{x}}$$

is a linear transformation of $\mathbf{x}$ with $\text{rank}(\mathbf{A}) = m \leq n$, then $\mathbf{y} \sim N_m(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$. We say that a vector random variable $\mathbf{x}$ follows a *singular normal distribution* if the covariance matrix $\boldsymbol{\Sigma}$ of $\mathbf{x}$ is singular, but if a maximal linearly independent subset of $\mathbf{x}$ is multivariately normally distributed.

A *bivariate-normal* density function for $\mu_1 = 5$, $\mu_2 = 6$, $\sigma_1 = 1.5$, $\sigma_2 = 3$, and $\rho_{12} = .5$ [i.e., $\sigma_{12} = (.5)(1.5)(3) = 2.25$] is depicted in Figure 3.11.

3.3.6 The Exponential Distributions

The exponential distributions are a continuous family indexed by the *rate parameter* λ, with density function

$$p(x) = \lambda e^{-\lambda x} \text{ for } x \geq 0$$

The expectation and variance of X are $E(X) = 1/\lambda$ and $V(X) = 1/\lambda^2$. Several exponential distributions for different values of the rate parameter appear in Figure 3.12. The exponential distributions, which are highly positively skewed, are frequently used to model time-to-event data when the "hazard" of occurrence of an event is constant during the period of observation.

[17] The converse is *not* true: Each X_i can be *univariately* normally distributed without $\mathbf{x}$ being *multivariate* normal.

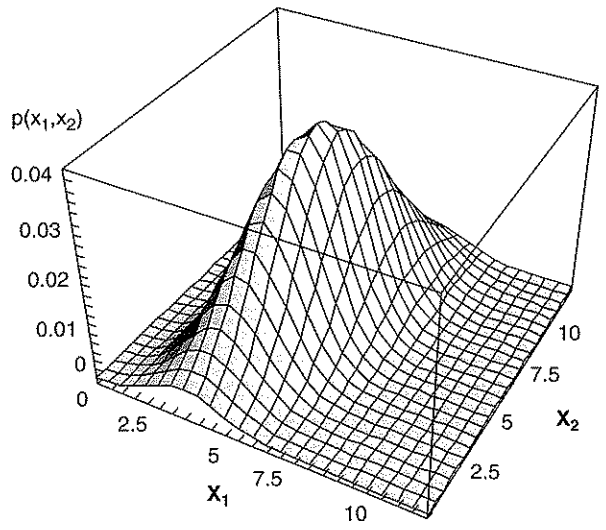

Figure 3.11 The bivariate-normal density function for $\mu_1 = 5$, $\mu_2 = 6$, $\sigma_1 = 1.5$, $\sigma_2 = 3$, and $\sigma_{12} = 2.25$. The slices of the density surface (representing the conditional distributions of each variable given values of the other) are normal both in the direction of X_1 and in the direction of X_2.

3.3.7 The Inverse-Gaussian Distributions

The inverse-Gaussian distributions are a continuous family indexed by two parameters, μ and λ, with density function

$$p(x) = \sqrt{\frac{\lambda}{2\pi x^3}} \exp\left[-\frac{\lambda(x-\mu)^2}{2x\mu^2}\right] \text{ for } x > 0$$

The expectation and variance of X are $E(X) = \mu$ and $V(X) = \mu^3/\lambda$. Figure 3.13 shows several inverse-Gaussian distributions. The variance of the inverse-Gaussian distribution increases with its mean; skewness also increases with the value of μ and decreases with λ.

The inverse-Gaussian distributions and the gamma distributions (immediately below) are often useful for modeling non-negative continuous data.

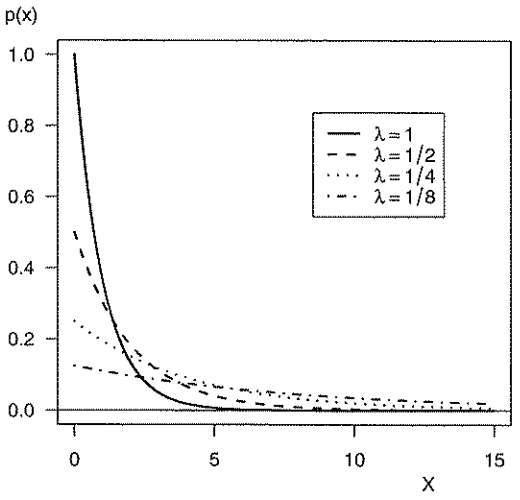

Figure 3.12 Exponential distributions for several values of the rate parameter λ.

3.3.8 The Gamma Distributions

The gamma distributions are a family of continuous distributions with probability density function indexed by the *scale parameter* $\omega > 0$ and *shape parameter* $\psi > 0$:

$$p(x) = \left(\frac{x}{\omega}\right)^{\psi-1} \frac{\exp\left(\frac{-x}{\omega}\right)}{\omega\Gamma(\psi)} \text{ for } x > 0$$

where $\Gamma(\cdot)$ is the gamma function (Equation 3.8 on page 105). The expectation and variance of the gamma distribution are, respectively, $E(X) = \omega\psi$ and $V(X) = \omega^2\psi$. Figure 3.14 shows gamma distributions for scale $\omega = 1$ and several values of the shape ψ. (Altering the scale parameter would change only the labeling of the horizontal axis in the graph.) As the shape parameter gets larger, the distribution grows more symmetric.

If $X_1, X_2, \ldots, X_k$ are independent gamma random variables with common scale ω and shape parameters $\psi_1, \psi_2, \ldots, \psi_k$, consecutively, then $X \equiv X_1 + X_2 + \cdots + X_k$ is gamma distributed with scale ω and shape $\psi = \psi_1 + \psi_2 + \cdots + \psi_k$.

The chi-squared distribution with n degrees of freedom is equal to the gamma distribution with scale parameter $\omega = 2$ and shape $\psi = n/2$. The

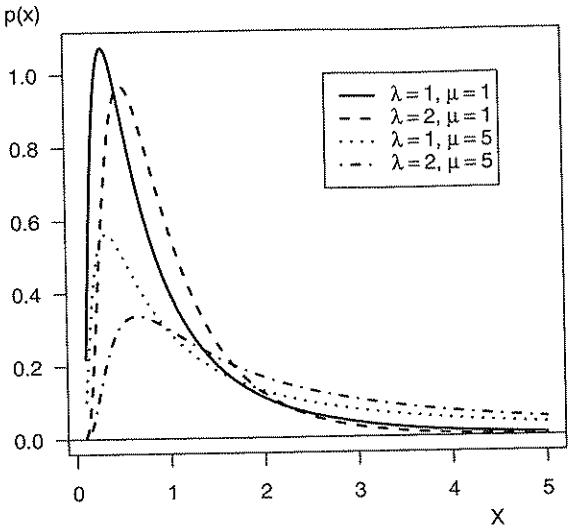

Figure 3.13 Inverse-Gaussian distributions for several combinations of values of the parameters μ and λ.

exponential distribution with rate parameter λ is equal to the the gamma distribution with shape $\psi = 1$ and scale $\omega = 1/\lambda$.

3.3.9 The Beta Distributions

The beta distributions are a family of continuous distributions with two *shape parameters* $\alpha > 0$ and $\beta > 0$, and with density function

$$p(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)} \text{ for } 0 \leq x \leq 1$$

where

$$B(\alpha, \beta) \equiv \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

is the *beta function*. The expectation and variance of the beta distribution are $E(X) = \alpha/(\alpha+\beta)$ and

$$V(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

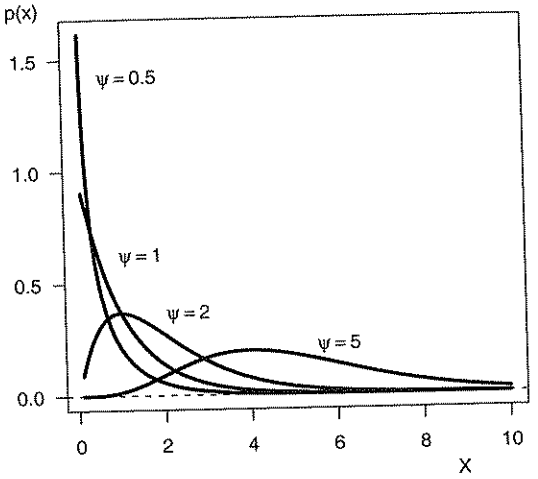

Figure 3.14 Several gamma distributions for scale $\omega = 1$ and various values of the shape parameter ψ.

The expectation, therefore, depends on the relative size of the parameters, with $E(X) = 0.5$ when $\alpha = \beta$. The skewness of the beta distribution also depends on the relative sizes of the parameters, and the distribution is symmetric when $\alpha = \beta$. The variance declines as α and β grow. Figure 3.15 shows several beta distributions. As is apparent from these graphs, the shape of the beta distribution is very flexible.

3.4 Asymptotic Distribution Theory: An Introduction

Partly because it is at times difficult to determine the small-sample properties of statistical estimators, it is of interest to investigate how an estimator behaves as the sample size grows. *Asymptotic distribution theory* provides tools for this investigation. I will merely outline the theory here: More complete accounts are available in many sources, including some of the references at the end of this chapter.

3.4.1 Probability Limits

Although asymptotic distribution theory applies to sequences of random variables, it is necessary first to consider the *nonstochastic infinite sequence* $\{a_1, a_2, \ldots, a_n, \ldots\}$. By "nonstochastic" I mean that each a_n is a fixed

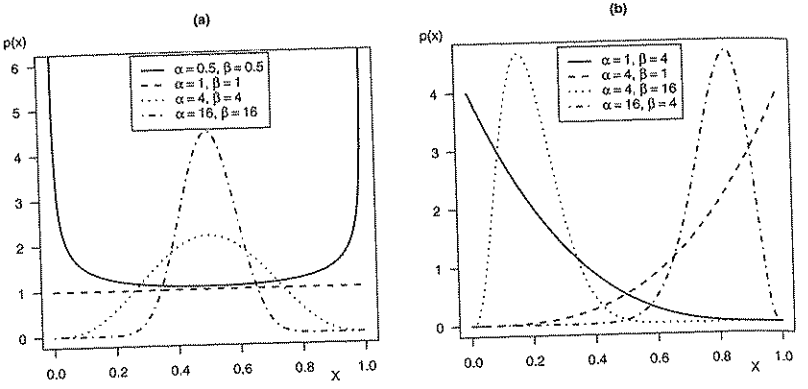

Figure 3.15 Beta distributions for several combinations of values of the shape parameters α and β. As is apparent in panel (a), the beta distribution reduces to the rectangular distribution when $\alpha = \beta = 1$. Symmetric beta distributions are shown in panel (a) and asymmetric distributions in panel (b).

number rather than a random variable. As the reader may be aware, this sequence has a *limit* a when, given any positive number ε, no matter how small, there is a positive integer $n(\varepsilon)$ such that $|a_n - a| < \varepsilon$ for all $n > n(\varepsilon)$. In words: a_n can be made arbitrarily close to a by picking n sufficiently large. The notation $n(\varepsilon)$ stresses that the required value of n depends on the selected criterion ε. (Cf., the definition of the limit of a *function*, discussed in Section 2.2.) To describe this state of affairs compactly, we write $\lim_{n\to\infty} a_n = a$. If, for example, $a_n = 1 + 1/n$, then $\lim_{n\to\infty} a_n = 1$; this sequence and its limit are graphed in Figure 3.16.

Consider now a *sequence of random variables* $\{X_1, X_2, \ldots, X_n, \ldots\}$. In a typical statistical application, X is some estimator and n is the size of the sample from which the estimator is calculated. Let $p_n \equiv \Pr(|X_n - a| < \delta)$, where a is a constant and δ is a small positive number. Think of p_n as the probability that X_n is *close* to a. Suppose that the *nonstochastic* sequence of probabilities $\{p_1, p_2, \ldots, p_n, \ldots\}$ approaches a limit of 1;[18] that is, $\lim_{n\to\infty} \Pr(|X_n - a| < \delta) = 1$. Then, as n grows, the random variable X_n

[18] To say that $\{p_1, p_2, \ldots, p_n, \ldots\}$ is a *nonstochastic* sequence is only apparently contradictory: Although these probabilities are based on random variables, the probabilities themselves are each specific numbers—such as, .6, .9, and so forth.

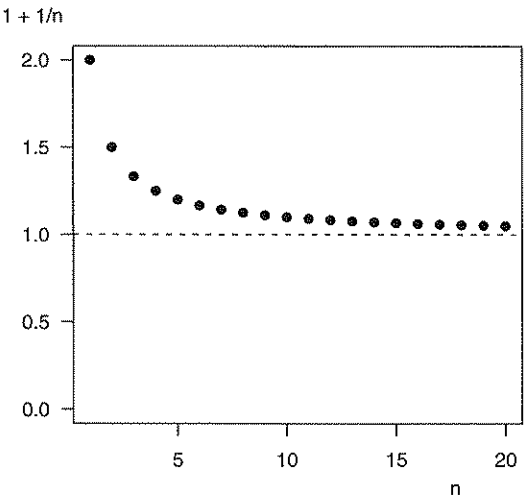

Figure 3.16 The first 20 values of the sequence $a_n = 1 + 1/n$, which has the limit $a = 1$.

concentrates more and more of its probability in a small region around a, a situation that is illustrated in Figure 3.17. If this result holds regardless of how small δ is, then we say that a is the *probability limit* of X_n, denoted plim $X_n = a$. It is common to drop the subscript n to write the even more compact expression, plim $X = a$.

Probability limits have the following very useful property: If plim $X = a$, and if $Y = f(X)$ is some continuous function of X, then plim $Y = f(a)$. Likewise, if plim $X = a$, plim $Y = b$, and $Z = f(X, Y)$ is a continuous function of X and Y, then plim $Z = f(a, b)$.

3.4.2 Asymptotic Expectation and Variance

We return to the sequence of random variables $\{X_1, X_2, \ldots, X_n, \ldots\}$. Let μ_n denote the expectation of X_n. Then $\{\mu_1, \mu_2, \ldots, \mu_n, \ldots\}$ is a non-stochastic sequence. If this sequence approaches a limit μ, then we call μ the *asymptotic expectation* of X, also written $\mathcal{E}(X)$.

Although it seems natural to define an asymptotic variance analogously as the limit of the sequence of variances, this definition is not satisfactory because (as the following example illustrates) $\lim_{n\to\infty} V(X_n)$ is 0 in most

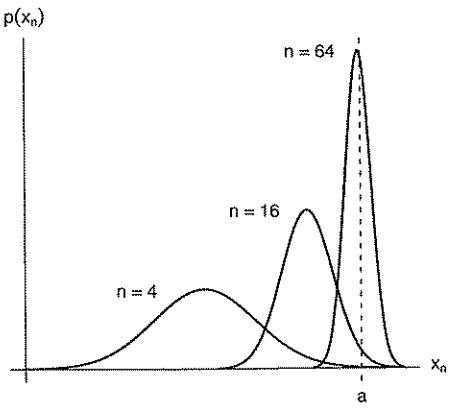

Figure 3.17 plim $X_n = a$: As n grows, the distribution of X_n concentrates more and more of its probability in a small region around a.

interesting cases. Suppose that we calculate the mean $\overline{X}_n$ for a sample of size n drawn from a population with mean μ and variance σ^2. We know, from elementary statistics, that $E(\overline{X}_n) = \mu$ and that

$$V(\overline{X}_n) = E[(\overline{X}_n - \mu)^2] = \frac{\sigma^2}{n}$$

Consequently, $\lim_{n\to\infty} V(\overline{X}_n) = 0$. Inserting the factor $\sqrt{n}$ within the square, however, produces the expectation $E\{[\sqrt{n}(\overline{X}_n - \mu)]^2\} = \sigma^2$. Dividing by n and taking the limit yields the answer that we want, defining the *asymptotic variance* of the sample mean:

$$\begin{aligned}\mathcal{V}(\overline{X}) &\equiv \lim_{n\to\infty} \frac{1}{n} E\{[\sqrt{n}(\overline{X}_n - \mu)]^2\} \\ &= \frac{1}{n}\mathcal{E}\{[\sqrt{n}(\overline{X}_n - \mu)]^2\} \\ &= \frac{\sigma^2}{n}\end{aligned}$$

This result is uninteresting for the present illustration because $\mathcal{V}(\overline{X}) = V(\overline{X})$; indeed, it is this equivalence that motivated the definition of the asymptotic variance in the first place. In certain applications, however, it

is possible to find the asymptotic variance of a statistic when the finite-sample variance is intractable. Then we can apply the asymptotic result as an approximation in large samples.

In the general case, where X_n has expectation μ_n, the asymptotic variance of X is defined to be[19]

$$\mathcal{V}(X) \equiv \frac{1}{n}\mathcal{E}\{[\sqrt{n}(X_n - \mu_n)]^2\} \tag{3.11}$$

3.4.3 Asymptotic Distribution

Let $\{P_1, P_2, \ldots, P_n, \ldots\}$ represent the CDFs of a sequence of random variables $\{X_1, X_2, \ldots, X_n, \ldots\}$. The CDF of X converges to the *asymptotic distribution* P if, given any positive number ε, however small, we can find a sufficiently large $n(\varepsilon)$ such that $|P_n(x) - P(x)| < \varepsilon$ for all $n > n(\varepsilon)$ and for all values x of the random variable.

A familiar illustration is provided by the *central-limit theorem*, which (in one of its versions) states that the mean of a set of independent and identically distributed random variables with finite expectations and variances follows an approximate normal distribution, the approximation improving as the number of random variables increases. Consider, for example, the mean of a sample of size n from the highly skewed exponential distribution with rate parameter $\lambda = 1$, for which the mean μ and variance σ^2 are both equal to 1. The exponential distribution is a special case of the gamma distribution, with shape parameter $\psi = 1$ and scale $\omega = 1/\lambda$; therefore, the sample sum $\sum_{i=1}^{n} X_i$ (and hence $n\overline{X}$) is gamma distributed with scale $\omega = 1$ and shape $\psi = n$. (The exponential and gamma distributions are described in Sections 3.3.6 and 3.3.8, respectively.) Figure 3.18 shows how the density function for the sampling distribution of the sample mean from this exponential population changes as the sample size grows, in each case comparing the true gamma sampling distribution of $\overline{X}$ with the normal approximation $N(1, 1/n)$: The normal approximation increases in accuracy (and the variance of the sampling distribution of $\overline{X}$ decreases) as the sample size gets larger.

[19] It is generally preferable to define asymptotic expectation and variance in terms of the asymptotic distribution (see the next section), because the sequences used for this purpose here do not exist in all cases (see Theil, 1971, pp. 375–376; also see McCallum, 1973). My use of the symbols $\mathcal{E}(\cdot)$ and $\mathcal{V}(\cdot)$ for asymptotic expectation and variance is not standard: The reader should be aware that these symbols are sometimes used in place of $E(\cdot)$ and $V(\cdot)$ to denote *ordinary* expectation and variance.

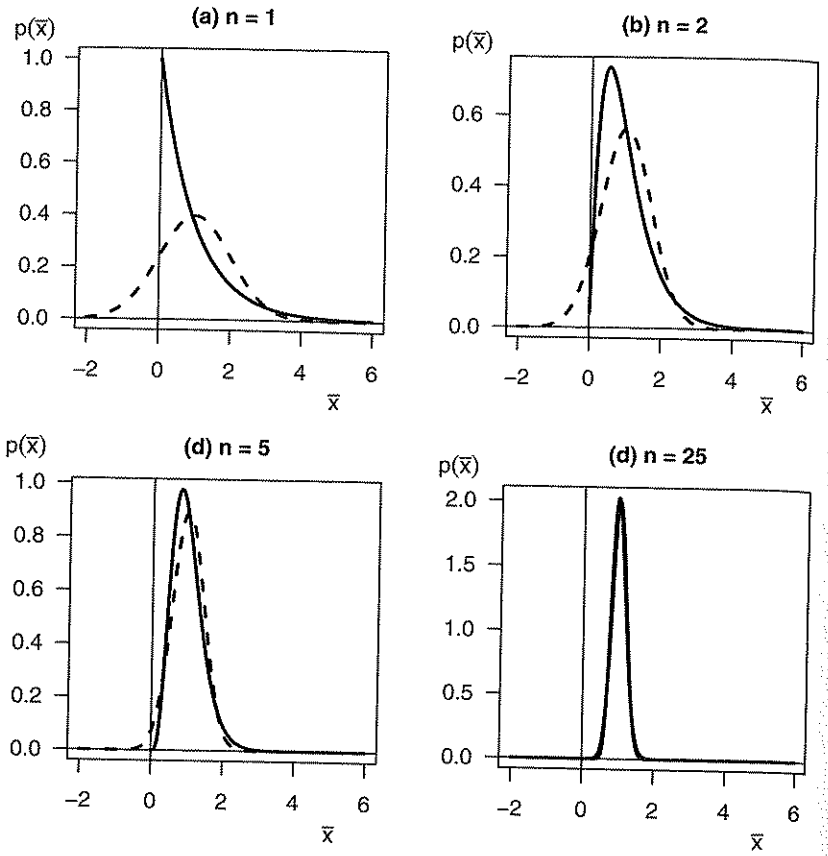

Figure 3.18 Illustration of the central limit theorem: The sampling distribution of the sample mean $\overline{X}$ for samples from an exponential population with rate parameter $\lambda = 1$, for various sample sizes n. Panel (a), for $n = 1$, corresponds to the population distribution of X. In each panel, the solid line gives the density function of the true (gamma) sampling distribution of $\overline{X}$, while the broken line gives the density function for the normal approximation $N(1, 1/n)$.

3.4.4 Vector and Matrix Random Variables

The results of this section extend straightforwardly to vectors and matrices: We say that $\underset{(m\times 1)}{\mathbf{x}} = \underset{(m\times 1)}{\mathbf{a}}$ when $\operatorname{plim} X_i = a_i$ for $i = 1, 2, \ldots, m$.

Likewise, plim $\underset{(m\times p)}{\mathbf{X}} = \underset{(m\times p)}{\mathbf{A}}$ means that plim $X_{ij} = a_{ij}$ for all i and j. The asymptotic expectation of the vector random variable $\underset{(m\times 1)}{\mathbf{x}}$ is defined as the vector of asymptotic expectations of its elements, $\boldsymbol{\mu} = \mathcal{E}(\mathbf{x}) \equiv [\mathcal{E}(X_1), \mathcal{E}(X_2), \ldots, \mathcal{E}(X_m)]'$. The asymptotic variance-covariance matrix of $\mathbf{x}$ is given by

$$\mathcal{V}(\mathbf{x}) \equiv \frac{1}{n}\mathcal{E}\{[\sqrt{n}(\mathbf{x}_n - \boldsymbol{\mu}_n)][\sqrt{n}(\mathbf{x}_n - \boldsymbol{\mu}_n)]'\}$$

3.5 Properties of Estimators[20]

An *estimator* is a sample statistic (i.e., a function of the observations of a sample) used to estimate an unknown population parameter. Because its value varies from one sample to the next, an estimator is a random variable. An *estimate* is the value of an estimator for a particular sample. The probability distribution of an estimator is called its *sampling distribution*; and the variance of this distribution is called the *sampling variance* of the estimator.

3.5.1 Bias

An estimator A of the parameter α is *unbiased* if $E(A) = \alpha$. The difference $E(A) - \alpha$ (which, of course, is 0 for an unbiased estimator) is the *bias* of A.

Suppose, for example, that we draw n independent observations X_i from a population with mean μ and variance σ^2. Then the sample mean $\overline{X} \equiv \sum X_i/n$ is an unbiased estimator of μ, while

$$S_*^2 \equiv \frac{\sum(X_i - \overline{X})^2}{n} \tag{3.12}$$

is a biased estimator of σ^2, because $E(S_*^2) = [(n-1)/n]\sigma^2$; the bias of S_*^2 is, therefore, $-\sigma^2/n$. Sampling distributions of unbiased and biased estimators are illustrated in Figure 3.19.

Asymptotic Bias The *asymptotic bias* of an estimator A of α is $\mathcal{E}(A) - \alpha$, and the estimator is *asymptotically unbiased* if $\mathcal{E}(A) = \alpha$. Thus, S_*^2 is asymptotically unbiased, because its bias $-\sigma^2/n \to 0$ as $n \to \infty$.

[20]Most of the material in this and the following section can be traced to a remarkable seminal paper on estimation by Fisher (1922)—arguably the most important statistical paper of the 20th century (see Aldrich, 1997).

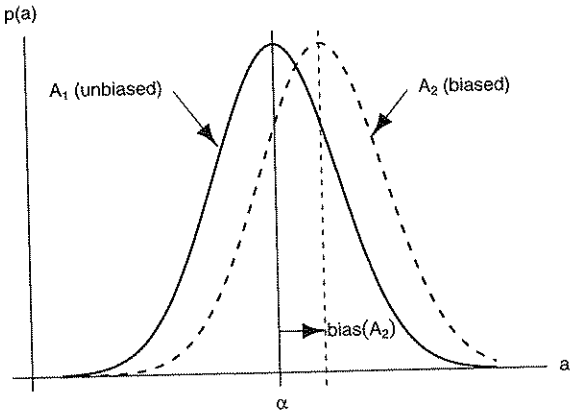

Figure 3.19 The estimator A_1 is an unbiased estimator of α because $E(A_1) = \alpha$; the estimator A_2 has a positive bias, because $E(A_2) > \alpha$.

3.5.2 Mean-Squared Error and Efficiency

To say that an estimator is unbiased means that its average value over repeated samples is equal to the parameter being estimated. This is clearly a desirable property for an estimator to possess, but it is cold comfort if the estimator does not provide estimates that are close to the parameter: In forming the expectation, large negative estimation errors for some samples could offset large positive errors for others.

The *mean-squared error* (*MSE*) of an estimator A of the parameter α is literally the average squared difference between the estimator and the parameter: $\text{MSE}(A) \equiv E[(A - \alpha)^2]$. The *efficiency* of an estimator is inversely proportional to its mean-squared error. We generally prefer a more efficient estimator to a less efficient one.

The mean-squared error of an unbiased estimator is simply its sampling variance, because $E(A) = \alpha$. For a biased estimator,

$$\begin{aligned}
\text{MSE}(A) &= E[(A - \alpha)^2] \\
&= E\{[A - E(A) + E(A) - \alpha]^2\} \\
&= E\{[A - E(A)]^2\} + [E(A) - \alpha]^2 + 2[E(A) - E(A)][E(A) - \alpha] \\
&= V(A) + [\text{bias}(A)]^2 + 0
\end{aligned}$$

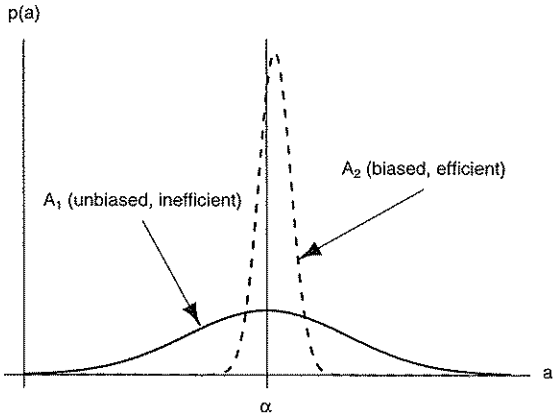

Figure 3.20 Relative efficiency of estimators: Even though it is biased, A_2 is a more efficient estimator of α than the unbiased estimator A_1, because the smaller variance of A_2 more than compensates for its small bias.

The efficiency of an estimator increases, therefore, as its sampling variance and bias decline. In comparing two estimators, an advantage in sampling variance can more than offset a disadvantage due to bias, as illustrated in Figure 3.20.

Asymptotic Efficiency *Asymptotic efficiency* is inversely proportional to *asymptotic mean-squared error (AMSE)* which, in turn, is the sum of asymptotic variance and squared asymptotic bias.

3.5.3 Consistency

An estimator A of the parameter α is *consistent* if $\operatorname{plim} A = \alpha$. A sufficient (but not necessary[21]) condition for consistency is that an estimator is asymptotically unbiased and that the sampling variance of the estimator approaches 0 as n increases; this condition implies that the mean-squared error of the estimator approaches a limit of 0. Figure 3.17 (page 116) illustrates consistency, if we construe X as an estimator of a.

[21] There are cases in which $\operatorname{plim} A = \alpha$, but the variance and asymptotic expectation of A do not exist. See Johnston (1972, p. 272) for an example.

The estimator S_*^2 given in Equation 3.12 (on page 119) is a consistent estimator of the population variance σ^2 even though it is biased in finite samples.

3.5.4 Sufficiency

Sufficiency is a more abstract property than unbias, efficiency, or consistency: A statistic S based on a sample of observations is *sufficient* for the parameter α if the statistic exhausts all of the information about α that is present in the sample. More formally, suppose that the observations X_1, $X_2, \ldots, X_n$ are drawn from a probability distribution with parameter α, and let the statistic $S \equiv f(X_1, X_2, \ldots, X_n)$. Then S is a sufficient statistic for α if the probability distribution of the observations *conditional* on the value of S, that is, $p(x_1, x_2, \ldots, x_n | S = s)$, does not depend on α. The sufficient statistic S *need not* be an estimator of α.

To illustrate the idea of sufficiency, suppose that n observations are independently sampled, and that each observation X_i takes on the value 1 with probability π and the value 0 with probability $1 - \pi$. That is, the X_i are independent, identically distributed Bernoulli random variables (see Section 3.2.1). I will demonstrate that the sample sum $S \equiv \sum_{i=1}^{n} X_i$ is a sufficient statistic for π: If we know the value s of S, then there are $\binom{n}{s}$ different possible arrangements of the s 1s and $n - s$ 0s, each with probability $1/\binom{n}{s}$. Note that the random variable S has a binomial distribution (see Section 3.2.1). Because this probability does not depend on the parameter π, the statistic S is sufficient for π. By a similar argument, the sample proportion $P \equiv S/n$ is also a sufficient statistic. The proportion P—but not the sum S—is an estimator of π.

The concept of sufficiency can be extended to sets of parameters and statistics: Given a sample of (possibly multivariate) observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, a vector of statistics $\mathbf{s} = [S_1, S_2, \ldots, S_p]' \equiv f(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ is *jointly sufficient* for the parameters $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_k]'$ if the conditional distribution of the observations given $\mathbf{s}$ does not depend on $\boldsymbol{\alpha}$. It can be shown, for example, that the mean $\overline{X}$ and variance S^2 calculated from an independent random sample are jointly sufficient statistics for the parameters μ and σ^2 of a normal distribution (as are the sample sum $\sum X_i$ and sum of squares $\sum X_i^2$, which jointly contain the same information as $\overline{X}$ and S^2). A set of sufficient statistics is called *minimally sufficient* if there is no smaller sufficient set.

3.5.5 Robustness

An estimator is said to be *robust* when its efficiency (and its efficiency relative to other estimators) does not strongly depend on the distribution of the data.

There is also another sense of robustness, termed *robustness of validity*, which is to be distinguished from *robustness of efficiency*. A procedure for statistical inference is said to be robust if its validity does not strongly depend on the distribution of the data. Thus the p-value for a robust hypothesis test is approximately correct even when the distributional assumptions (e.g., an assumption of normality) on which the test is based are violated. Similarly, a confidence interval is robust if it has approximately the stated level of coverage—for example, a 95% confidence interval covers the parameter in roughly 95% of samples—even when distributional assumptions are violated. Robustness of validity can be cold comfort when a test or confidence interval is based on an inefficient estimator: The test may have low power and the confidence interval may be very wide.

To make robustness of efficiency more concrete, let us focus on a simple setting: estimating the center μ of a symmetric distribution.[22] As long as X has finite variance σ^2, the variance of the sample mean $\overline{X}$ is $V(\overline{X}) = \sigma^2/n$, where n is the size of the sample (a familiar result from basic statistics). The variance of the sample median, however, depends on the distribution of X:

$$V(\text{median}) \approx \frac{1}{4n[p(x_{.5})]^2}$$

where $p(x_{.5})$ is the density at the (population) median of X.

Applied to a normal population, $X \sim N(\mu, \sigma^2)$, the variance of the median is $V(\text{median}) = \pi\sigma^2/2n$, and therefore the sample median is a less efficient estimator of μ than the sample mean is:

$$\frac{V(\text{median})}{V(\overline{X})} = \frac{\pi\sigma^2/2n}{\sigma^2/n} = \frac{\pi}{2} \approx 1.57$$

We would need a sample more than one-and-a-half times as large to estimate μ with a specified degree of precision using the sample median rather than the mean.

[22] In the *absence* of symmetry, what we mean by the center of the distribution becomes ambiguous.

In contrast, suppose now that X is t-distributed with 3 degrees of freedom, a distribution with heavier tails than the normal distribution. Then (using the properties of the t-distribution given in Section 3.3.3), $\sigma^2 = 3/(3-2) = 3$, $p(x_{.5}) = p(0) = 0.3675$, and, consequently,

$$V(\overline{X}) = \frac{3}{n}$$

$$V(\text{median}) = \frac{1}{4n(0.3675^2)} = \frac{1.851}{n}$$

In this case, therefore, the mean is only $1.851/3 = 0.617$ (i.e., 62%) as efficient as the median.

Robustness is closely related to *resistance* to unusual data: A resistant estimator is little affected by a small fraction of outlying data. The mean has low resistance to outliers, as is simply demonstrated: I drew a sample of six observations from the standard-normal distribution, obtaining

$$\begin{array}{lll} X_1 = -0.068 & X_2 = -1.282 & X_3 = 0.013 \\ X_4 = 0.141 & X_5 = -0.980 & X_6 = 1.263 \end{array} \tag{3.13}$$

The mean of these six values is $\overline{X} = -0.152$. Now imagine adding a seventh observation, X_7, allowing it to take on all possible values from -10 to $+10$ (or, with greater imagination, from $-\infty$ to $+\infty$). The result, called the *influence function* of the mean, is graphed in Figure 3.21(a). It is apparent from this figure that as the discrepant seventh observation grows more extreme, the sample mean chases it.

A related concept in assessing resistance is the *breakdown point* of an estimator: The breakdown point is the fraction of "bad" data that the estimator can tolerate without being affected to an arbitrarily large extent. The mean has a breakdown point of 0, because, as we have seen, a *single* bad observation can change the mean by an arbitrary amount. The median, in contrast, has a breakdown point of 50%, because fully half the data can be bad without causing the median to become completely unstuck.

***M*-Estimation** The mean minimizes the least-squares *objective function*

$$\sum_{i=1}^{n} \rho_{\text{LS}}(X_i - \widehat{\mu}) \equiv \sum_{i=1}^{n} (X_i - \widehat{\mu})^2$$

The shape of the influence function for the mean follows from the derivative of the objective function with respect to the *residual* $E \equiv X - \widehat{\mu}$:

$$\psi_{\text{LS}}(E) \equiv \rho'_{\text{LS}}(E) = 2E$$

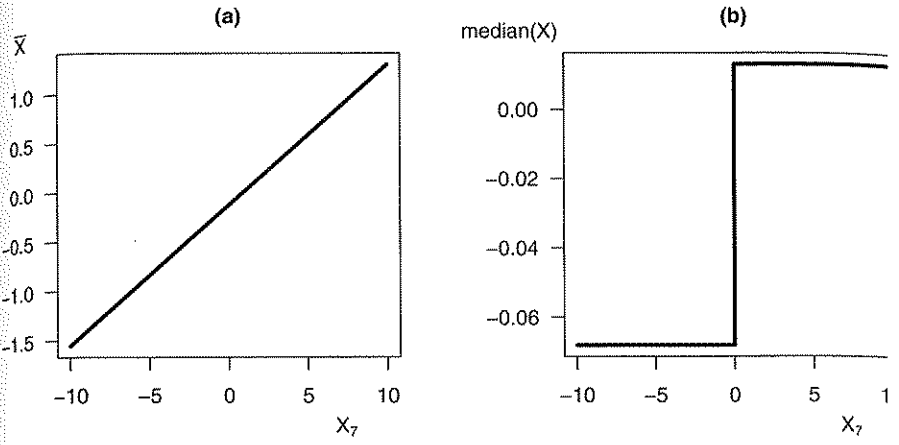

Figure 3.21 The influence functions for the mean (a) and median (b) for the sample $X_1 = -0.068, X_2 = -1.282, X_3 = 0.013, X_4 = 0.141, X_5 = -0.980, X_6 = 1.263$. The influence function for the median is bounded, while that for the mean is not. Note that the vertical axes for the two graphs have different scales.

Influence, therefore, is proportional to E. It is convenient to redefine the least-squares objective function as $\rho_{\text{LS}}(E) \equiv \frac{1}{2}E^2$, so that $\psi_{\text{LS}}(E) = E$.

Now consider the sample median as an estimator of μ. The median minimizes the *least-absolute-values* (LAV) objective function:

$$\sum_{i=1}^{n} \rho_{\text{LAV}}(E_i) = \sum_{i=1}^{n} \rho_{\text{LAV}}(X_i - \widehat{\mu}) \equiv \sum_{i=1}^{n} |X_i - \widehat{\mu}|$$

As a result, the median is much more resistant than the mean to outliers. The influence function of the median for the illustrative sample is shown in Figure 3.21(b). In contrast to the mean, the influence of a discrepant observation on the median is *bounded*. Once again, the derivative of the objective function gives the shape of the influence function:[23]

$$\psi_{\text{LAV}}(E) \equiv \rho'_{\text{LAV}}(E) = \begin{cases} 1 & \text{for } E > 0 \\ 0 & \text{for } E = 0 \\ -1 & \text{for } E < 0 \end{cases}$$

[23] Strictly speaking, the derivative of ρ_{LAV} is undefined at $E = 0$, but setting $\psi_{\text{LAV}}(0) \equiv 0$ is convenient.

Although the median is more resistant than the mean to outliers, we have seen that it is less efficient than the mean if the distribution of X is normal. Other objective functions combine resistance to outliers with greater robustness of efficiency. Estimators that can be expressed as minimizing an objective function $\sum_{i=1}^{n} \rho(E_i)$ are called *M-estimators*.[24]

Two common M-estimators are the *Huber* and the *biweight* (or *bisquare*). The Huber estimator is named after Peter J. Huber, who introduced M-estimation; the biweight is due to John W. Tukey, who made many important contributions to statistics, including to robust estimation.

- The Huber objective function is a compromise between least squares and least absolute values, behaving like least squares near the center of the data and like least absolute values in the tails:

$$\rho_H(E) \equiv \begin{cases} \frac{1}{2}E^2 & \text{for } |E| \leq k \\ k|E| - \frac{1}{2}k^2 & \text{for } |E| > k \end{cases}$$

The Huber objective function ρ_H and its derivative, the influence function ψ_H, are graphed in Figure 3.22:[25]

$$\psi_H(E) = \begin{cases} k & \text{for } E > k \\ E & \text{for } |E| \leq k \\ -k & \text{for } E < -k \end{cases}$$

The value k, which defines the center and tails, is called a *tuning constant*. It is most natural to express the tuning constant as a multiple of the *scale* (i.e., the spread) of the variable X, that is, to take $k = cS$, where S is a measure of scale. The sample standard deviation is a poor measure of scale in this context, because it is even more affected than the mean by outliers. A common robust measure of scale is the *median absolute deviation* (MAD):

$$\text{MAD} \equiv \text{median}|X_i - \widehat{\mu}|$$

The estimate $\widehat{\mu}$ can be taken, at least initially, as the median value of X. We can then define $S \equiv \text{MAD}/0.6745$, which ensures that S estimates the standard deviation σ when the population is normal. Using $k = 1.345S$

[24] Estimators that can be written in this form can be thought of as generalizations of maximum-likelihood estimators (see Section 3.6), hence the term "M"-estimator. The maximum-likelihood estimator is produced by taking $\rho_{\text{ML}}(x - \mu) \equiv -\log_e p(x - \mu)$ for an appropriate probability or probability-density function $p(\cdot)$.

[25] My terminology here is loose, but convenient: Strictly speaking, the ψ-function is not the influence function, but it has the same shape as the influence function.

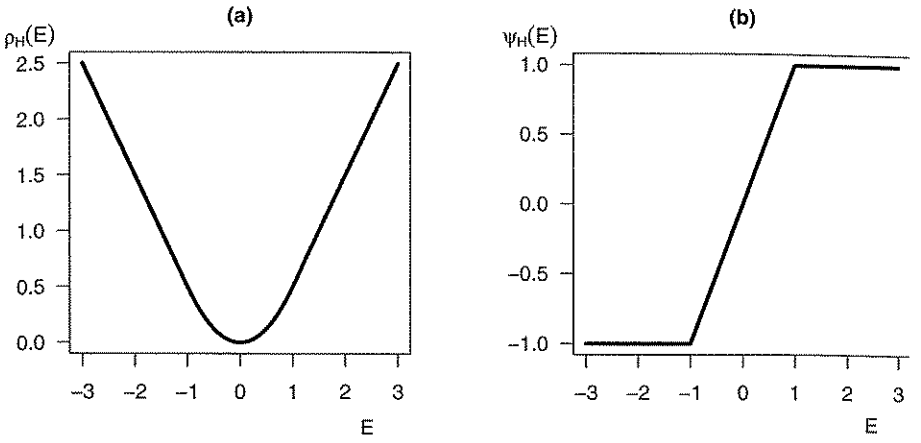

Figure 3.22 Huber objective function ρ_H (a) and "influence function" ψ_H (b). To calibrate these graphs, the tuning constant is set to $k = 1$. (See the text for a discussion of the tuning constant.)

(i.e., $1.345/0.6745 \approx 2$ MADs) produces 95% efficiency relative to the sample mean when the population is normal, along with considerable resistance to outliers when it is not. A smaller tuning constant can be employed for more resistance.

- The biweight objective function levels off at very large residuals:[26]

$$\rho_{\text{BW}}(E) \equiv \begin{cases} \frac{k^2}{6}\left\{1 - \left[1 - \left(\frac{E}{k}\right)^2\right]^3\right\} & \text{for } |E| \leq k \\ \frac{k^2}{6} & \text{for } |E| > k \end{cases}$$

The influence function for the biweight estimator, therefore, "re-descends" to 0, *completely discounting* observations that are sufficiently outlying:

$$\psi_{\text{BW}}(E) = \begin{cases} E\left[1 - \left(\frac{E}{k}\right)^2\right]^2 & \text{for } |E| \leq k \\ 0 & \text{for } |E| > k \end{cases}$$

The functions ρ_{BW} and ψ_{BW} are graphed in Figure 3.23. Using $k = 4.685S$ (i.e., $4.685/0.6745 \approx 7$ MADs) produces 95% efficiency when sampling from a normal population.

[26] The term "bisquare" applies literally to the ψ-function and to the weight function (hence "biweight"), to be introduced presently—not to the objective function.

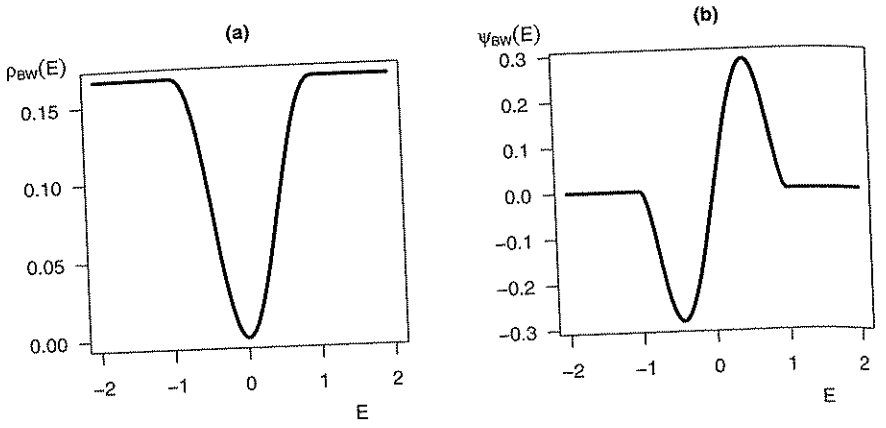

Figure 3.23 Biweight objective function ρ_{BW} (a) and "influence function" ψ_{BW} (b). To calibrate these graphs, the tuning constant is set to $k = 1$. The influence function "redescends" to 0 when $|E|$ is large.

Both the Huber and the biweight estimators achieve a breakdown point of 50% when the MAD is used to estimate scale.

Calculation of M-estimators usually requires an *iterative* (i.e., repetitive) procedure (although iteration is not necessary for the mean and median which, as we have seen, fit into the M-estimation framework). An estimating equation for $\widehat{\mu}$ is obtained by setting the derivative of the objective function (with respect to $\widehat{\mu}$) to 0, obtaining

$$\sum_{i=1}^{n} \psi\left(X_i - \widehat{\mu}\right) = 0 \tag{3.14}$$

There are several general approaches to solving Equation 3.14; probably the most straightforward, and the simplest to implement computationally, is to re-weight the mean iteratively:

1. Define the *weight function* $w(E) \equiv \psi(E)/E$. Then the estimating equation becomes

$$\sum_{i=1}^{n} \left(X_i - \widehat{\mu}\right) w_i = 0 \tag{3.15}$$

TABLE 3.1
Weight Functions $w(E) = \psi(E)/E$ for Several M-Estimators

Objective Function $\rho(E)$	*Weight Function* $w(E)$
Least squares	1
Least absolute values	$1/\lvert E\rvert$ (for $E \neq 0$)
Huber	1 for $\lvert E\rvert \leq k$ $k/\lvert E\rvert$ for $\lvert E\rvert > k$
Bisquare (biweight)	$\left[1 - \left(\frac{E}{k}\right)^2\right]^2$ for $\lvert E\rvert \leq k$ 0 for $\lvert E\rvert > k$

where

$$w_i \equiv w(X_i - \widehat{\mu})$$

The solution of Equation 3.15 is the weighted mean,

$$\widehat{\mu} = \frac{\sum w_i X_i}{\sum w_i}$$

The weight functions corresponding to the least-squares, LAV, Huber, and bisquare objective functions are shown in Table 3.1 and graphed in Figure 3.24. The least-squares weight function accords equal weight to each observation, while the bisquare gives 0 weight to observations that are sufficiently outlying; the LAV and Huber weight functions descend toward 0 but never quite reach it.

2. Select an initial estimate $\widehat{\mu}^{(0)}$, such as the median of the X-values.[27] Using $\widehat{\mu}^{(0)}$, calculate an initial estimate of scale $S^{(0)}$ and initial weights $w_i^{(0)} = w(X_i - \widehat{\mu}^{(0)})$. Set the iteration counter $l = 0$. The scale is required to calculate the tuning constant $k = cS$ (for prespecified c).
3. At each iteration l, calculate $\widehat{\mu}^{(l)} = \sum w_i^{(l-1)} X_i / \sum w_i^{(l-1)}$. Stop when the change in $\widehat{\mu}^{(l)}$ is negligible from one iteration to the next.

To illustrate the application of these estimators, recall our sample of six observations from the standard normal distribution $N(0, 1)$ (given in Equation 3.13 on page 124); let us contaminate the sample with the outlying

[27] Because the estimating equation for a re-descending M-estimator, such as the bisquare, can have more than one root, the selection of an initial estimate might be consequential.

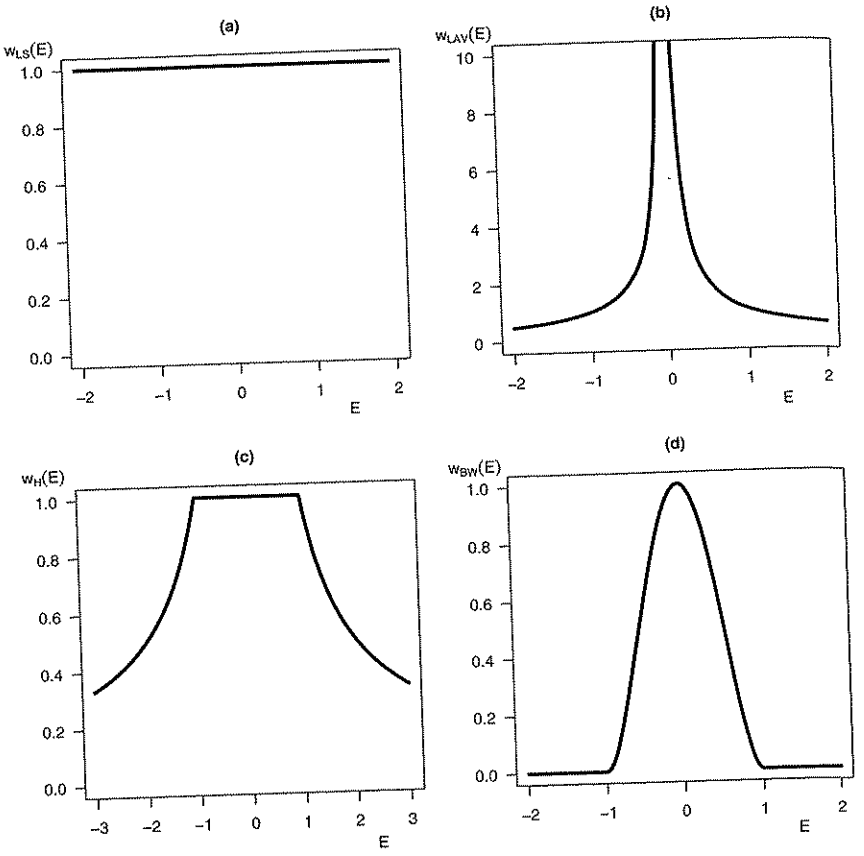

Figure 3.24 Weight functions $w(E)$ for the (a) least-squares, (b) least-absolute-values, (c) Huber, and (d) biweight estimators. The tuning constants for the Huber and biweight estimators are taken as $k = 1$. Note that the vertical axis in the graph for the LAV estimator and the horizontal axis in the graph for the Huber estimator are different from the others.

value $X_7 = 10$. Using $c = 1.345$ for the Huber estimator and $c = 4.685$ for the biweight,

$$\overline{X} = 1.298, \text{median}(X) = 0.013, \widehat{\mu}_H = 0.201, \text{and } \widehat{\mu}_{BW} = -0.161$$

It is clear that the sample mean $\overline{X}$ is seriously affected by the outlier but that the other estimators are not.

3.6 Maximum-Likelihood Estimation

The *method of maximum likelihood* provides estimators that have both a reasonable intuitive basis and many desirable statistical properties. The method is very broadly applicable and is simple to apply. Moreover, once a maximum-likelihood estimator is derived, the general theory of maximum-likelihood estimation provides standard errors, statistical tests, and other results useful for statistical inference. A disadvantage of the method, however, is that it frequently requires strong assumptions about the structure of the data.

The likelihood function plays a central role in classical statistical inference, but it is also important in Bayesian inference (discussed in Section 3.7).

3.6.1 Preliminary Example

Let us first consider a simple example: Suppose that we want to estimate the probability π of getting a head on flipping a particular coin. We flip the coin independently 10 times (i.e., we sample $n = 10$ flips), obtaining the following result: $HHTHHHTTHH$. The probability of obtaining this sequence—in advance of collecting the data—is a function of the unknown parameter π:

$$\begin{aligned}\Pr(\text{data}|\text{parameter}) &= \Pr(HHTHHHTTHH|\pi)\\ &= \pi\pi(1-\pi)\pi\pi\pi(1-\pi)(1-\pi)\pi\pi\\ &= \pi^7(1-\pi)^3\end{aligned}$$

This is simply the product of probabilities for 10 independent Bernoulli random variables (taking $X_i = 1$ for a head and $X_i = 0$ for a tail, $i = 1, \ldots, 10$; see Section 3.2.1).

The data for our particular sample are *fixed*, however: We have already collected them. The parameter π also has a fixed value, but this value is unknown, and so we can let it vary in our imagination between 0 and 1, treating the probability of the observed data as a function of π. This function is called the *likelihood function*:

$$\begin{aligned}L(\text{parameter}|\text{data}) &= L(\pi|HHTHHHTTHH)\\ &= \pi^7(1-\pi)^3\end{aligned}$$

The probability function and the likelihood function are the same equation, but the probability function is a function of the data with the value of the parameter fixed, while the likelihood function is a function of the parameter with the data fixed.

Here are some representative values of the likelihood for different values of π:[28]

π	$L(\pi\|\text{data}) = \pi^7(1-\pi)^3$
0.0	0.0
.1	.0000000729
.2	.00000655
.3	.0000750
.4	.000354
.5	.000977
.6	.00179
.7	.00222
.8	.00168
.9	.000478
1.0	0.0

The full likelihood function is graphed in Figure 3.25. Although each value of $L(\pi|\text{data})$ is a notional probability, the function $L(\pi|\text{data})$ is *not* a probability distribution or a density function: It does not integrate to 1, for example.

In the present instance, the probability of obtaining the sample of data that we have in hand, *HHTHHHTTHH*, is small regardless of the true value of π. This is usually the case: Unless the sample is very small, *any specific* sample result—including the one that is realized—will have low probability in advance of collecting data.

Nevertheless, the likelihood contains useful information about the unknown parameter π. For example, π *cannot* be 0 or 1, because if it were either of these values, then the observed data (which includes both heads and tails) could not have been obtained. Reversing this reasoning, the value of π that is most supported by the data is the one for which the likelihood is

[28]The likelihood is a *continuous* function of π for values of π between 0 and 1. This contrasts, in the present case, with the probability function, because there is a *finite* number $\left(2^{10} = 1024\right)$ of possible samples.

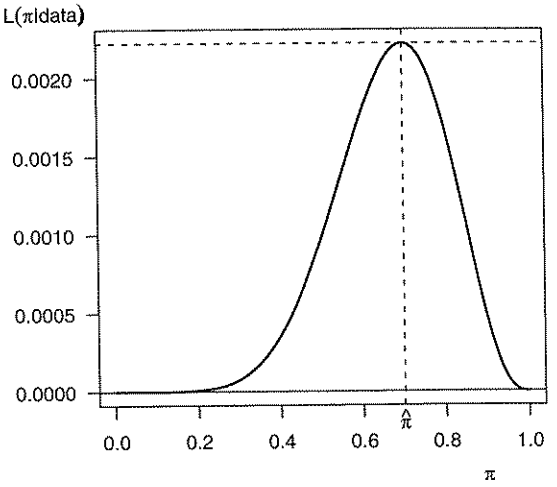

Figure 3.25 The likelihood function $L(\pi|HHTHHHTTHH) = \pi^7(1-\pi)^3$.

largest. This value is the *maximum-likelihood estimate* (*MLE*), denoted $\widehat{\pi}$. Here, $\widehat{\pi} = .7$, which is just the sample proportion of heads, 7/10.

Generalization of the Example More generally, for n independent flips of the coin, producing a particular sequence that includes x heads and $n - x$ tails,

$$L(\pi|\text{data}) = \Pr(\text{data}|\pi) = \pi^x(1-\pi)^{n-x}$$

We want the value of π that maximizes $L(\pi|\text{data})$, which we often abbreviate $L(\pi)$. As is typically the case, it is simpler—and equivalent—to find the value of π that maximizes the *log of the likelihood*, here

$$\log_e L(\pi) = x \log_e \pi + (n-x)\log_e(1-\pi) \tag{3.16}$$

Differentiating $\log_e L(\pi)$ with respect to π,

$$\begin{aligned}\frac{d\log_e L(\pi)}{d\pi} &= \frac{x}{\pi} + (n-x)\frac{1}{1-\pi}(-1) \\ &= \frac{x}{\pi} - \frac{n-x}{1-\pi}\end{aligned}$$

The derivative of the log likelihood with respect to the parameter is called the *score* (or *score function*). Setting the score to 0 and solving for π produces the MLE which, as before, is the sample proportion x/n (as the reader may wish to verify). The maximum-likelihood *estimator* is $\widehat{\pi} = X/n$. To avoid this slightly awkward substitution of estimator for estimate in the last step, we can replace x by X in the log likelihood function (Equation 3.16).

3.6.2 Properties of Maximum-Likelihood Estimators

Under very broad conditions, maximum-likelihood estimators have the following general properties (see the references at the end of this chapter):

- Maximum-likelihood estimators are consistent.
- They are asymptotically unbiased, although they may be biased in finite samples.
- They are asymptotically efficient—no asymptotically unbiased estimator has a smaller asymptotic variance.
- They are asymptotically normally distributed.
- If there is a sufficient statistic for a parameter, then the maximum-likelihood estimator of the parameter is a function of a sufficient statistic.
- If $\widehat{\alpha}$ is the MLE of α, and if $\beta = f(\alpha)$ is a function of α, then $\widehat{\beta} = f(\widehat{\alpha})$ is the MLE of β.
- The asymptotic sampling variance of the MLE $\widehat{\alpha}$ of a parameter α can be obtained from the second derivative of the log likelihood:

$$V(\widehat{\alpha}) = \frac{1}{-E\left[\frac{d^2 \log_e L(\alpha)}{d\alpha^2}\right]} \tag{3.17}$$

The denominator of $V(\widehat{\alpha})$ is called the *expected or Fisher information*,[29]

$$\mathcal{I}(\alpha) \equiv -E\left[\frac{d^2 \log_e L(\alpha)}{d\alpha^2}\right]$$

[29] Strictly speaking, the Fisher information is the variance of the score evaluated at the parameter value α:

$$\mathcal{I}(\alpha) = E\left[\left(\frac{d \log_e L(\alpha)}{d\alpha}\right)^2 \middle| \alpha\right]$$

In most instances, however, this is equivalent to the generally more convenient formula given in the text. Notice that the variance of the score is simply the expectation of its square, because the expected score is 0 at α.

In practice, we substitute the MLE $\widehat{\alpha}$ into Equation 3.17 to obtain an *estimate* of the asymptotic sampling variance, $\widehat{V}(\widehat{\alpha})$.[30]

- $L(\widehat{\alpha})$ is the value of the likelihood function at the MLE $\widehat{\alpha}$, while $L(\alpha)$ is the likelihood for the true (but generally unknown) parameter α. The *log-likelihood-ratio statistic*

$$G^2 \equiv 2\log_e \frac{L(\widehat{\alpha})}{L(\alpha)} = 2[\log_e L(\widehat{\alpha}) - \log_e L(\alpha)]$$

follows an asymptotic chi-square distribution with 1 degree of freedom. Because, by definition, the MLE maximizes the likelihood for our *particular* sample, the value of the likelihood at the true parameter value α is generally *smaller* than at the MLE $\widehat{\alpha}$ (unless, by good fortune, $\widehat{\alpha}$ and α happen to coincide).

Establishing these results is well beyond the scope of this chapter, but the results do make some intuitive sense. For example, if the log likelihood has a sharp peak, then the MLE is clearly differentiated from nearby values. Under these circumstances, the second derivative of the log likelihood is a large negative number; there is a lot of "information" in the data concerning the value of the parameter; and the sampling variance of the MLE is small. If, in contrast, the log likelihood is relatively flat at its maximum, then alternative estimates quite different from the MLE are nearly as good as the MLE; there is little information in the data concerning the value of the parameter; and the sampling variance of the MLE is large (see Figure 3.26).

3.6.3 Statistical Inference: Wald, Likelihood-Ratio, and Score Tests

The properties of maximum-likelihood estimators described in the previous section lead directly to three common and general procedures—called the *Wald test*, the *likelihood-ratio test*, and the *score test*—for testing the statistical hypothesis H_0: $\alpha = \alpha_0$. The score test is sometimes called the *Lagrange-multiplier test*. (Lagrange multipliers are described in Section 2.5.2.) The Wald and likelihood-ratio tests can be "turned around" to produce confidence intervals for α.

[30] It is also possible, and sometimes computationally advantageous, to base an estimate of the variance of the MLE $\widehat{\alpha}$ on the *observed information*,

$$\mathcal{I}_O(\widehat{\alpha}) \equiv \frac{d^2 \log_e L(\widehat{\alpha})}{d\widehat{\alpha}^2}$$

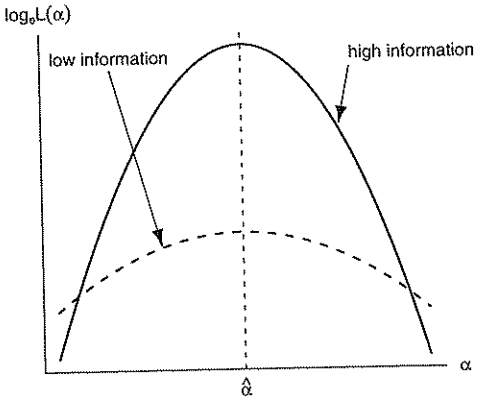

Figure 3.26 Two imagined log likelihoods: one strongly peaked, providing high information about the the parameter α; and the other flat, providing low information about α.

- *Wald test.* Relying on the asymptotic normality of the MLE $\widehat{\alpha}$, we calculate the test statistic

$$Z_0 \equiv \frac{\widehat{\alpha} - \alpha_0}{\sqrt{\widehat{\mathcal{V}}(\widehat{\alpha})}}$$

  which is asymptotically distributed as $N(0, 1)$ under H_0.
- *Likelihood-ratio test.* Employing the log-likelihood ratio, the test statistic

$$G_0^2 \equiv 2 \log_e \frac{L(\widehat{\alpha})}{L(\alpha_0)} = 2[\log_e L(\widehat{\alpha}) - \log_e L(\alpha_0)]$$

  is asymptotically distributed as χ_1^2 under H_0.
- *Score test.* Recall that the score $S(\alpha) \equiv d \log_e L(\alpha)/d\alpha$ is the slope of the log likelihood at a particular value of α. At the MLE, the score is 0: $S(\widehat{\alpha}) = 0$. It can be shown that the *score statistic*

$$S_0 \equiv \frac{S(\alpha_0)}{\sqrt{\mathcal{I}(\alpha_0)}}$$

  is asymptotically distributed as $N(0, 1)$ under H_0.

Unless the log likelihood is quadratic, the three test statistics can produce somewhat different results in specific samples, although the tests are asymptotically equivalent. In certain contexts, the score test has the practical advantage of not requiring the computation of the MLE $\widehat{\alpha}$ (because

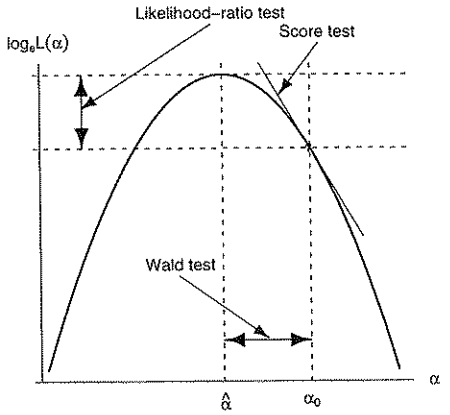

Figure 3.27 Tests of the hypothesis H_0: $\alpha = \alpha_0$: The likelihood-ratio test compares $\log_e L(\widehat{\alpha})$ to $\log_e L(\alpha_0)$; the Wald test compares $\widehat{\alpha}$ to α_0; and the score test examines the slope of $\log_e L(\alpha)$ at $\alpha = \alpha_0$.

S_0 depends only on the null value α_0, which is specified in H_0). In most instances, however, the likelihood-ratio test is more reliable than the Wald and score tests in smaller samples.

Figure 3.27 shows the relationship among the three test statistics, and clarifies the intuitive rationale of each: The Wald test measures the distance between $\widehat{\alpha}$ and α_0, using the standard error to calibrate this distance. If $\widehat{\alpha}$ is far from α_0, for example, then doubt is cast on H_0. The likelihood-ratio test measures the distance between $\log_e L(\widehat{\alpha})$ and $\log_e L(\alpha_0)$; if $\log_e L(\widehat{\alpha})$ is much larger than $\log_e L(\alpha_0)$, then H_0 is probably wrong. Finally, the score test statistic measures the slope of log likelihood at α_0; if this slope is very steep, then we are probably far from the peak of the likelihood function, casting doubt on H_0.

An Illustration It is instructive to apply these results to our previous example, in which we sought to estimate the probability π of obtaining a head from a coin based on a sample of n independent flips. Recall that the MLE of π is the sample proportion $\widehat{\pi} = X/n$, where X counts the number of heads in the sample. The second derivative of the log likelihood (Equation 3.16 on page 133) is

$$\frac{d^2 \log_e L(\pi)}{d\pi^2} = -\frac{X}{\pi^2} - \left[-\frac{n-X}{(1-\pi)^2}(-1)\right]$$
$$= \frac{-X + 2\pi X - n\pi^2}{\pi^2(1-\pi)^2}$$

Noting that $E(X) = n\pi$, the expected information is

$$\mathcal{I}(\pi) = \frac{-n\pi + 2n\pi^2 - n\pi^2}{-\pi^2(1-\pi^2)} = \frac{n}{\pi(1-\pi)}$$

and the asymptotic variance of $\widehat{\pi}$ is $\mathcal{V}(\widehat{\pi}) = [\mathcal{I}(\pi)]^{-1} = \pi(1-\pi)/n$, a familiar result: In this case, the asymptotic variance coincides with the exact, finite-sample variance of $\widehat{\pi}$. The *estimated* asymptotic sampling variance is $\widehat{\mathcal{V}}(\widehat{\pi}) = \widehat{\pi}(1-\widehat{\pi})/n$.

For our sample of $n = 10$ flips with $X = 7$ heads, $\widehat{\mathcal{V}}(\widehat{\pi}) = (.7 \times .3)/10 = 0.0210$, and a 95% asymptotic confidence interval for π based on the Wald statistic is

$$\pi = .7 \pm 1.96 \times \sqrt{0.0210} = .7 \pm .284$$

where, recall, 1.96 is the standard-normal value with probability .025 to the right. Alternatively, to test the hypothesis that the coin is fair, H_0: $\pi = .5$, we can calculate the Wald test statistic

$$Z_0 = \frac{.7 - .5}{\sqrt{0.0210}} = 1.38$$

which corresponds to a two-tail p-value [from $N(0, 1)$] of .168.

The log likelihood, recall, is

$$\log_e L(\pi) = X \log_e \pi + (n - X) \log_e (1 - \pi)$$
$$= 7 \log_e \pi + 3 \log_e (1 - \pi)$$

Using this equation,

$$\log_e L(\widehat{\pi}) = \log_e L(.7) = 7 \log_e .7 + 3 \log_e .3 = -6.1086$$
$$\log_e L(\pi_0) = \log_e L(.5) = 7 \log_e .5 + 3 \log_e .5 = -6.9315$$

The likelihood-ratio test statistic for H_0 is, therefore,

$$G_0^2 = 2[-6.1086 - (-6.9315)] = 1.646$$

which corresponds to a p-value (from χ_1^2) of .199.

Finally, for the score test,

$$S(\pi) = \frac{d \log_e L(\pi)}{d\pi} = \frac{X}{\pi} - \frac{n - X}{1 - \pi}$$

from which

$$S(\pi_0) = \frac{7}{.5} - \frac{3}{.5} = 8$$

Evaluating the expected information at π_0 produces

$$\mathcal{I}(\pi_0) = \mathcal{I}(.5) = \frac{10}{.5 \times .5} = 40$$

The score statistic is, therefore,

$$S_0 = \frac{S(\pi_0)}{\sqrt{\mathcal{I}(\pi_0)}} = \frac{8}{\sqrt{40}} = 1.265$$

for which the two-tail p-value [from $N(0, 1)$] is .206.

The three tests are in reasonable agreement, but all are quite inaccurate! An exact test, using the null binomial distribution of X (the number of heads: see Section 3.2.1),

$$p(x) = \binom{10}{x} .5^x .5^{10-x} = \binom{10}{x} .5^{10}$$

yields a two-tail p-value of .3438 [corresponding to $\Pr(X \leq 3 \text{ or } X \geq 7)$]. The lesson to be drawn from this example is that we must be careful in applying asymptotic results to small samples.

3.6.4 Several Parameters

The maximum-likelihood method can be generalized to simultaneous estimation of several parameters. Let $p(\underset{(n\times m)}{\mathbf{X}} \mid \underset{(k\times 1)}{\boldsymbol{\alpha}})$ represent the probability or probability density for n possibly multivariate observations $\mathbf{X}$ ($m \geq 1$), which depend on k independent parameters $\boldsymbol{\alpha}$.[31] The likelihood

[31] To say that the parameters are *independent* means that the value of none can be obtained from the values of the others. If there is a dependency among the parameters, then the redundant parameter can simply be replaced by a function of other parameters.

$L(\boldsymbol{\alpha}) \equiv L(\boldsymbol{\alpha}|\mathbf{X})$ is a function of the parameters $\boldsymbol{\alpha}$, and we seek the values $\widehat{\boldsymbol{\alpha}}$ that maximize this function. As before, it is generally more convenient to work with $\log_e L(\boldsymbol{\alpha})$ in place of $L(\boldsymbol{\alpha})$. To maximize the likelihood, we find the vector partial derivative $\partial \log_e L(\boldsymbol{\alpha})/\partial \boldsymbol{\alpha}$, set this derivative to $\mathbf{0}$, and solve the resulting matrix equation for $\widehat{\boldsymbol{\alpha}}$. If there is more than one root, then we choose the solution that produces the largest likelihood.

As in the case of a single parameter, the maximum-likelihood estimator is consistent, asymptotically unbiased, asymptotically efficient, asymptotically normal (but now *multivariate* normal), and based on sufficient statistics. The asymptotic variance-covariance matrix of the MLE is

$$\underset{(k\times k)}{\mathcal{V}(\widehat{\boldsymbol{\alpha}})} = \left\{-E\left[\frac{\partial^2 \log_e L(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}\, \partial \boldsymbol{\alpha}'}\right]\right\}^{-1} \tag{3.18}$$

The matrix in braces in Equation 3.18 is called the *expected* or *Fisher information matrix*, $\mathcal{I}(\boldsymbol{\alpha})$ (not to be confused with the identity matrix $\mathbf{I}$).[32] Moreover, if $\boldsymbol{\beta} = f(\boldsymbol{\alpha})$, then the MLE of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}} = f(\widehat{\boldsymbol{\alpha}})$. Notice how the formulas for several parameters closely parallel those for one parameter.

Generalizations of the score and Wald tests follow directly. The Wald statistic for H_0: $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0$ is

$$Z_0^2 \equiv (\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)'\widehat{\mathcal{V}}(\widehat{\boldsymbol{\alpha}})^{-1}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)$$

The score vector is $S(\boldsymbol{\alpha}) \equiv \partial \log_e L(\boldsymbol{\alpha})/\partial \boldsymbol{\alpha}$; and the score statistic is

$$S_0^2 \equiv S(\boldsymbol{\alpha}_0)'\mathcal{I}(\boldsymbol{\alpha}_0)^{-1}S(\boldsymbol{\alpha}_0)$$

The likelihood-ratio test also generalizes straightforwardly:

$$G_0^2 \equiv 2\log_e\left[\frac{L(\widehat{\boldsymbol{\alpha}})}{L(\boldsymbol{\alpha}_0)}\right]$$

All three test statistics are asymptotically distributed as χ_k^2 under H_0.

---

[32] As in the case of a single parameter, a slightly more general definition of the Fisher information is

$$\mathcal{I}(\boldsymbol{\alpha}) = E\left[\left(\frac{\partial \log_e L(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\right)^2\Big|_{\boldsymbol{\alpha}}\right]$$

Similarly, it is also possible to work with the *observed* information at the MLE $\widehat{\boldsymbol{\alpha}}$.

Each of these tests can be adapted to more complex hypotheses. Suppose, for example, that we wish to test the hypothesis H_0 that p of the k elements of $\boldsymbol{\alpha}$ are equal to particular values. Let $L(\widehat{\boldsymbol{\alpha}}_0)$ represent the maximized likelihood under the constraint represented by the hypothesis (i.e., setting the p parameters equal to their hypothesized values, but leaving the other parameters free to be estimated); $L(\widehat{\boldsymbol{\alpha}})$ represents the globally maximized likelihood when the constraint is relaxed. Then, under the hypothesis H_0,

$$G_0^2 \equiv 2 \log_e \left[\frac{L(\widehat{\boldsymbol{\alpha}})}{L(\widehat{\boldsymbol{\alpha}}_0)} \right]$$

has an asymptotic chi-square distribution with p degrees of freedom.

The following example (adapted from Theil, 1971, pp. 389–390) illustrates these results: A sample of n independent observations X_i is drawn from a normally distributed population with unknown mean μ and variance σ^2. We want to estimate μ and σ^2. The likelihood function is

$$\begin{aligned} L(\mu, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(X_i - \mu)^2}{2\sigma^2} \right] \\ &= (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2 \right] \end{aligned}$$

and the log likelihood is

$$\log_e L(\mu, \sigma^2) = -\frac{n}{2} \log_e 2\pi - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum (X_i - \mu)^2$$

with partial derivatives

$$\begin{aligned} \frac{\partial \log_e L(\mu, \sigma^2)}{\partial \mu} &= \frac{1}{\sigma^2} \sum (X_i - \mu) \\ \frac{\log_e L(\mu, \sigma^2)}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum (X_i - \mu)^2 \end{aligned}$$

Setting the partial derivatives to 0 and solving simultaneously for the maximum-likelihood estimators of μ and σ^2 produces

$$\begin{aligned} \widehat{\mu} &= \frac{\sum X_i}{n} = \overline{X} \\ \widehat{\sigma^2} &= \frac{\sum (X_i - \overline{X})^2}{n} \end{aligned}$$

The matrix of second partial derivatives of the log likelihood is

$$\begin{bmatrix} \frac{\partial^2 \log_e L}{\partial \mu^2} & \frac{\partial^2 \log_e L}{\partial \mu\, \partial \sigma^2} \\ \frac{\partial^2 \log_e L}{\partial \sigma^2\, \partial \mu} & \frac{\partial^2 \log_e L}{\partial (\sigma^2)^2} \end{bmatrix} = \begin{bmatrix} -\frac{n}{\sigma^2} & -\frac{1}{\sigma^4}\sum(X_i - \mu) \\ -\frac{1}{\sigma^4}\sum(X_i - \mu) & \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum(X_i - \mu)^2 \end{bmatrix}$$

Taking expectations, noting that $E(X_i - \mu) = 0$ and that $E[(X_i - \mu)^2] = \sigma^2$, produces the negative of the expected information matrix:

$$-\mathcal{I}(\mu, \sigma^2) = \begin{bmatrix} -\frac{n}{\sigma^2} & 0 \\ 0 & -\frac{n}{2\sigma^4} \end{bmatrix}$$

The asymptotic variance-covariance matrix of the maximum-likelihood estimators is, as usual, the inverse of the information matrix:

$$\mathcal{V}(\widehat{\mu}, \widehat{\sigma}^2) = [\mathcal{I}(\mu, \sigma^2)]^{-1} = \begin{bmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix}$$

The result for the sampling variance of $\widehat{\mu} = \overline{X}$ is the usual one (σ^2/n). The MLE of σ^2 is biased but consistent (and, indeed, is the estimator S_*^2 given previously in Equation 3.12 on page 119).

In many applications, including the examples in this chapter, the data comprise an independent random sample of n identically distributed observations. The likelihood for the data is then the product of likelihood-components for the observations, $L_i(\boldsymbol{\alpha})$, and the log likelihood is the sum of the logs of these components:

$$\log_e L(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \log_e L_i(\boldsymbol{\alpha})$$

The score function consequently is also a sum of case-wise terms:

$$S(\boldsymbol{\alpha}) = \sum_{i=1}^{n} S_i(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \frac{\partial \log_e L_i(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}$$

Finally, the information for the sample is n times the information in an individual observation (denoted $\mathcal{I}_1$):

$$\mathcal{I}(\boldsymbol{\alpha}) = n\mathcal{I}_1(\boldsymbol{\alpha}) = nE\left[\frac{\partial^2 \log_e L_i(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}\, \partial \boldsymbol{\alpha}'}\right]$$

This last result holds because the expectation of the second partial derivative of the likelihood is identical for all n observations.

3.6.5 The Delta Method

As I have explained, if $\beta = f(\alpha)$, and if $\widehat{\alpha}$ is the maximum-likelihood estimator of α, then $\widehat{\beta} = f(\widehat{\alpha})$ is the maximum-likelihood estimator of β. This result implies that $\widehat{\beta}$ is asymptotically normally distributed with asymptotic expectation β, even when the function $f(\cdot)$ is nonlinear.

The *delta method* produces an estimate of the asymptotic variance of $\widehat{\beta}$ based on a first-order Taylor-series approximation (see Section 2.6) to $f(\widehat{\alpha})$ around the true value of the parameter α:

$$\widehat{\beta} = f(\widehat{\alpha}) \approx f(\alpha) + f'(\alpha)(\widehat{\alpha} - \alpha) \tag{3.19}$$

Here, $f'(\alpha) = df(\alpha)/d\alpha$ is the derivative of $f(\alpha)$ with respect to α.

The first term on the right-hand side of Equation 3.19, $f(\alpha)$, is a constant (because the parameter α has a fixed value), and the second term is linear in $\widehat{\alpha}$ [again because α, and hence $f'(\alpha)$, are constants]; thus

$$V(\widehat{\beta}) \approx \left[f'(\alpha)\right]^2 V(\widehat{\alpha})$$

where $V(\widehat{\alpha})$ is the asymptotic variance of $\widehat{\alpha}$. In practice, we substitute the MLE $\widehat{\alpha}$ for α to obtain the *estimated* asymptotic variance of $\widehat{\beta}$:

$$\widehat{V}(\widehat{\beta}) = \left[f'(\widehat{\alpha})\right]^2 V(\widehat{\alpha})$$

To illustrate the application of the delta method, recall that the sample proportion $\widehat{\pi}$ is the maximum-likelihood estimator of the population proportion π, with asymptotic (and, indeed, finite-sample) variance $V(\widehat{\pi}) = \pi(1-\pi)/n$, where n is the sample size. The *log-odds*, or *logit*, is defined as

$$\Lambda = f(\pi) \equiv \log_e \frac{\pi}{1-\pi}$$

The MLE of Λ is therefore $\widehat{\Lambda} = \log_e[\widehat{\pi}/(1-\widehat{\pi})]$, and the asymptotic sampling variance of the sample logit is

$$\begin{aligned} V(\widehat{\Lambda}) &\approx \left[f'(\pi)\right]^2 V(\widehat{\pi}) \\ &= \left[\frac{1}{\pi(1-\pi)}\right]^2 \frac{\pi(1-\pi)}{n} \\ &= \frac{1}{n\pi(1-\pi)} \end{aligned}$$

Finally, the estimated asymptotic sampling variance of the logit is $\widehat{\mathcal{V}}(\widehat{\Lambda}) = 1/[n\widehat{\pi}(1-\widehat{\pi})]$.

The delta method extends readily to functions of several parameters: Suppose that $\beta \equiv f(\alpha_1, \alpha_2, \ldots, \alpha_k) = f(\boldsymbol{\alpha})$, and that $\widehat{\boldsymbol{\alpha}}$ is the MLE of $\boldsymbol{\alpha}$, with asymptotic covariance matrix $\mathcal{V}(\widehat{\boldsymbol{\alpha}})$. Then the asymptotic variance of $\widehat{\beta} = f(\widehat{\boldsymbol{\alpha}})$ is

$$\mathcal{V}(\widehat{\beta}) \approx \left[\mathbf{g}(\boldsymbol{\alpha})\right]' \mathcal{V}(\widehat{\boldsymbol{\alpha}})\mathbf{g}(\boldsymbol{\alpha}) = \sum_{i=1}^{k}\sum_{j=1}^{k} v_{ij} \times \frac{\partial\widehat{\beta}}{\partial\alpha_i} \times \frac{\partial\widehat{\beta}}{\partial\alpha_j}$$

where $\mathbf{g}(\boldsymbol{\alpha}) \equiv \partial\widehat{\beta}/\partial\boldsymbol{\alpha}$ and v_{ij} is the i, jth entry of $\mathcal{V}(\widehat{\boldsymbol{\alpha}})$. The estimated asymptotic variance of $\widehat{\beta}$ is thus

$$\widehat{\mathcal{V}}(\widehat{\beta}) = \left[\mathbf{g}(\widehat{\boldsymbol{\alpha}})\right]' \mathcal{V}(\widehat{\boldsymbol{\alpha}})\mathbf{g}(\widehat{\boldsymbol{\alpha}})$$

The delta method is not only applicable to functions of maximum-likelihood estimators, but more generally to functions of estimators that are asymptotically normally distributed.

3.7 Introduction to Bayesian Inference

This section introduces Bayesian statistics, an alternative approach to statistical inference. The treatment here is very brief, presenting the principal ideas of Bayesian inference but not developing the topic in any detail.

3.7.1 Bayes's Theorem

Recall (from Section 3.1.1) the definition of *conditional probability*: The probability of an event A given that another event B is known to have occurred is

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} \tag{3.20}$$

Likewise, the conditional probability of B given A is

$$\Pr(B|A) = \frac{\Pr(A \cap B)}{\Pr(A)} \tag{3.21}$$

Solving Equation 3.21 for the *joint probability* of A and B produces

$$\Pr(A \cap B) = \Pr(B|A)\Pr(A)$$

and substituting this result into Equation 3.20 yields *Bayes's theorem*:

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)} \tag{3.22}$$

Bayes's theorem is named after its discoverer, the Reverend Thomas Bayes, an 18th-century English mathematician.

Bayesian statistical inference is based on the following interpretation of Equation 3.22: Let A represent some uncertain proposition whose truth or falsity we wish to establish—for example, the proposition that a parameter is equal to a particular value. Let B represent observed data that are relevant to the truth of the proposition. We interpret the unconditional probability $\Pr(A)$, called the *prior probability* of A, as our strength of belief in the truth of A prior to collecting data, and $\Pr(B|A)$ as the probability of obtaining the observed data assuming the truth of A—that is, the *likelihood* of the data given A (in the sense of the preceding section). The *unconditional* probability of the data B is[33]

$$\Pr(B) = \Pr(B|A)\Pr(A) + \Pr(B|\overline{A})\Pr(\overline{A})$$

Then $\Pr(A|B)$, given by Equation 3.22 and called the *posterior probability* of A, represents our revised strength of belief in A in light of the data B.

Bayesian inference is therefore a rational procedure for updating one's beliefs on the basis of evidence. This *subjectivist* interpretation of probabilities contrasts with the *frequentist* interpretation of probabilities as long-run proportions (see Section 3.1.1). Bayes's theorem follows from elementary probability theory *whether or not* one accepts its subjectivist interpretation, but it is the latter that gives rise to common procedures of Bayesian statistical inference.

Preliminary Example Consider the following simple (if contrived) example: Suppose that you are given a gift of two "biased" coins, one of which produces heads with probability $\Pr(H) = .3$ and the other with $\Pr(H) = .8$. Each of these coins comes in a box marked with its bias,

[33]This is an application of the *law of total probability*: Given an event B and a set of k disjoint events $A_1, \ldots, A_k$ for which $\sum_{i=1}^{k} \Pr(A_i) = 1$ (i.e., the events A_i partition the sample space S),

$$\Pr(B) = \sum_{i=1}^{k} \Pr(B|A_i)\Pr(A_i)$$

but you carelessly misplace the boxes and put the coins in a drawer; a year later, you do not remember which coin is which. To try to distinguish the coins, you pick one arbitrarily and flip it 10 times, obtaining the data *HHTHHHTTHH*—that is, a particular sequence of 7 heads and 3 tails. (These are the "data" used in a preliminary example of maximum-likelihood estimation in Section 3.6.1.)

Let A represent the event that the selected coin has $\Pr(H) = .3$; then $\overline{A}$ is the event that the coin has $\Pr(H) = .8$. Under these circumstances, it seems reasonable to take as prior probabilities $\Pr(A) = \Pr(\overline{A}) = .5$. The likelihood of the data under A and $\overline{A}$ is

$$\Pr(B|A) = .3^7(1 - .3)^3 = .0000750$$
$$\Pr(B|\overline{A}) = .8^7(1 - .8)^3 = .0016777$$

Notice that, as is typically the case, the likelihood of the observed data is small in both cases, but the data are much more likely under $\overline{A}$ than under A. (The likelihood of these data for *any* value of $\Pr(H)$ between 0 and 1 was shown previously in Figure 3.25 on page 133.)

Using Bayes's theorem (Equation 3.22), you find the posterior probabilities

$$\Pr(A|B) = \frac{.0000750 \times .5}{.0000750 \times .5 + .0016777 \times .5} = .0428$$
$$\Pr(\overline{A}|B) = \frac{.0016777 \times .5}{.0000750 \times .5 + .0016777 \times .5} = .9572$$

suggesting that it is much more probable that the selected coin has $\Pr(H) = .8$ than $\Pr(H) = .3$.

3.7.2 Extending Bayes's Theorem

Bayes's theorem extends readily to situations in which there are more than two hypotheses A and $\overline{A}$: Let the various hypotheses be represented by $H_1, H_2, \ldots, H_k$, with prior probabilities $\Pr(H_i)$, $i = 1, \ldots, k$ that sum to 1;[34] and let D represent the observed data, with likelihood $\Pr(D|H_i)$ under hypothesis H_i. Then the posterior probability of hypothesis H_i is

[34]To employ Bayesian inference, your prior beliefs must be consistent with probability theory, and so the prior probabilities must sum to 1.

$$\Pr(H_i|D) = \frac{\Pr(D|H_i)\Pr(H_i)}{\sum_{j=1}^{k}\Pr(D|H_j)\Pr(H_j)} \tag{3.23}$$

The denominator in Equation 3.23 insures that the posterior probabilities for the various hypotheses sum to 1. It is sometimes convenient to omit this normalization, simply noting that

$$\Pr(H_i|D) \propto \Pr(D|H_i)\Pr(H_i)$$

that is, that the posterior probability of a hypothesis is proportional to the product of the likelihood under the hypothesis and its prior probability. If necessary, we can always divide by $\sum \Pr(D|H_i)\Pr(H_i)$ to recover the posterior probabilities.

Bayes's theorem is also applicable to random variables: Let α represent a parameter of interest, with prior probability distribution or density $p(\alpha)$, and let $L(\alpha) \equiv p(D|\alpha)$ represent the likelihood function for the parameter α. Then

$$p(\alpha|D) = \frac{L(\alpha)p(\alpha)}{\sum_{\text{all }\alpha'} L(\alpha')p(\alpha')}$$

when the parameter α is discrete, or

$$p(\alpha|D) = \frac{L(\alpha)p(\alpha)}{\int L(\alpha')p(\alpha')\,d\alpha'}$$

when, as is more common, α is continuous. In either case,

$$p(\alpha|D) \propto L(\alpha)p(\alpha)$$

That is, the posterior distribution or density is proportional to the product of the likelihood and the prior distribution or density. As before, we can if necessary divide by $\sum L(\alpha)p(\alpha)$ or $\int L(\alpha)p(\alpha)\,d\alpha$ to recover the posterior probabilities or densities.

The following points are noteworthy:

- We require a prior distribution $p(\alpha)$ over the possible values of the parameter α (the *parameter space*) to set the machinery of Bayesian inference in motion.
- In contrast to classical statistics, we treat the parameter α as a *random variable* rather than as an unknown *constant*. We retain Greek letters for parameters, however, because in contrast to the data, parameters are never known with certainty—even after collecting data.

Conjugate Priors The mathematics of Bayesian inference is especially simple when the prior distribution is selected so that the likelihood and prior combine to produce a posterior distribution that is in the same family as the prior. In this case, we say that the prior distribution is a *conjugate prior*.

At one time, Bayesian inference was only practical when conjugate priors were employed, limiting its scope of application. Advances in computer software and hardware, however, make it practical to evaluate mathematically intractable posterior distributions by simulated random sampling. Such *Markov-chain Monte-Carlo* ("*MCMC*") methods have produced a flowering of Bayesian applied statistics. Nevertheless, the choice of prior distribution can be an important one.

3.7.3 An Example of Bayesian Inference

Continuing the previous example, suppose more realistically that you are given a coin and wish to estimate the probability π that the coin turns up heads, but cannot restrict π in advance to a small number of discrete values; rather, π could, in principle, be any number between 0 and 1. To estimate π, you plan to gather data by independently flipping the coin 10 times. We know from our previous work that the Bernoulli likelihood is

$$L(\pi) = \pi^h (1-\pi)^{10-h} \tag{3.24}$$

where h is the observed number of heads. You conduct the experiment, obtaining the data *HHTHHHTTHH*, and thus $h = 7$.

The conjugate prior for the Bernoulli likelihood in Equation 3.24 is the beta distribution (Section 3.3.9),

$$p(\pi) = \frac{\pi^{a-1}(1-\pi)^{b-1}}{B(a,b)} \text{ for } 0 \leq \pi \leq 1 \text{ and } a, b \geq 0$$

When you multiply the beta prior by the likelihood, you get a posterior density of the form

$$p(\pi|D) \propto \pi^{h+a-1}(1-\pi)^{10-h+b-1} = \pi^{6+a}(1-\pi)^{2+b}$$

that is, a beta distribution with shape parameters $h + a = 7 + a$ and $10 - h + b = 3 + b$. Put another way, the prior in effect adds a heads and b tails to the likelihood.

How should you select a and b? One approach would be to reflect your subjective assessment of the likely value of π. For example, you might examine the coin and note that it seems to be reasonably well balanced, suggesting that π is probably close to .5. Picking $a = b = 16$ would in effect confine your estimate of π to the range between .3 and .7 (see Figure 3.15 on page 114). If you are uncomfortable with this restriction, then you could select smaller values of a and b: When $a = b = 1$, all values of π are equally likely—a so-called *flat prior distribution*, reflecting complete ignorance about the value of π.[35]

Figure 3.28 shows the posterior distribution for π under these two priors. Under the flat prior, the posterior is proportional to the likelihood, and therefore if you take the mode of the posterior as your estimate of π, you get the MLE $\widehat{\pi} = .7$.[36] The *informative prior* $a = b = 16$, in contrast, has a mode at $\pi \approx .55$, which is much closer to the mode of the prior distribution $\pi = .5$.

It may be disconcerting that the conclusion should depend so crucially on the prior distribution, but this result is a product of the very small sample in the example: Recall that using a beta prior in this case is like adding $a + b$ observations to the data. As the sample size grows, the likelihood comes to dominate the posterior distribution, and the influence of the prior distribution fades.[37] In the current example, if the coin is flipped n times, then the posterior distribution takes the form

$$p(\pi|D) \propto \pi^{h+a-1}(1-\pi)^{n-h+b-1}$$

---

[35] In this case, the prior is a rectangular density function, with the parameter π bounded between 0 and 1. In other cases, such as estimating the mean μ of a normal distribution, which is unbounded, a flat prior of the form $p(\mu) = c$ (for any positive constant c) over $-\infty < \mu < \infty$ does not enclose a finite probability, and hence cannot represent a density function. When combined with the likelihood, such an *improper prior* can nevertheless lead to a proper posterior distribution—that is, to a posterior density that integrates to 1.

A more subtle point is that a flat prior for one parametrization of a probability model for the data need not be flat for an alternative parametrization: For example, suppose that you take the odds $\omega \equiv \pi/(1-\pi)$ as the parameter of interest, or the logit $\lambda \equiv \log_e[\pi/(1-\pi)]$; a flat prior for π is not flat for ω or for λ.

[36] An alternative is to take the *mean* of the posterior distribution as a point estimate of π. In most cases, however, the posterior distribution will approach a normal distribution as the sample size increases, and the mean and mode will therefore be approximately equal if the sample size is sufficiently large.

[37] An exception to this rule occurs when the prior distribution assigns zero density to some values of the parameter; such values will necessarily have posterior densities of zero as well.

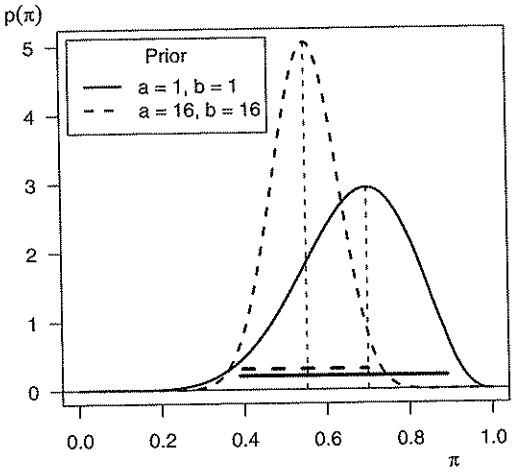

Figure 3.28 Posterior distributions of the probability of a head π under two prior distributions: The flat beta prior with $a = 1$, $b = 1$, and the informative beta prior with $a = 16, b = 16$. The data contain 7 heads in 10 flips of a coin. The two horizontal lines near the bottom of the graph show 95% central posterior intervals corresponding to the two priors.

and the numbers of heads h and tails $n - h$ will grow with the number of flips. It is intuitively sensible that your prior beliefs should carry greater weight when the sample is small than when it is large.

3.7.4 Bayesian Interval Estimates

As in classical inference, it is desirable not only to provide a point estimate of a parameter but also to quantify uncertainty in the estimate. The posterior distribution of the parameter displays statistical uncertainty in a direct form. From the posterior distribution, one can compute various kinds of Bayesian interval estimates, which are analogous to classical confidence intervals.

A very simple choice is the *central posterior interval*: The $100a$ percent central posterior interval runs from the $(1 - a)/2$ to the $(1 + a)/2$ quantile of the posterior distribution. Unlike a classical confidence interval, however, the interpretation of which is famously convoluted (to the confusion of innumerable students of basic statistics), a Bayesian posterior interval has

a simple interpretation as a probability statement: The probability is .95 that the parameter is in the 95% posterior interval. This difference reflects the Bayesian interpretation of a parameter as a random variable, with the posterior distribution expressing subjective uncertainty about the value of the parameter after observing the data.

Ninety-five percent central posterior intervals for the example are shown for the two posterior distributions in Figure 3.28.

3.7.5 Bayesian Inference for Several Parameters

Bayesian inference extends straightforwardly to the simultaneous estimation of several parameters $\boldsymbol{\alpha} \equiv [\alpha_1, \alpha_2, \ldots, \alpha_k]'$. In this case, it is necessary to specify a *joint prior distribution* for the parameters, $p(\boldsymbol{\alpha})$, along with the *joint likelihood*, $L(\boldsymbol{\alpha})$. Then, as in the case of one parameter, the *joint posterior distribution* is proportional to the product of the prior distribution and the likelihood:

$$p(\boldsymbol{\alpha}|D) \propto p(\boldsymbol{\alpha})L(\boldsymbol{\alpha})$$

Inference typically focuses on the *marginal posterior distribution* of each parameter, $p(\alpha_i|D)$.

3.8 Recommended Reading

Almost any introductory text in mathematical statistics, and many econometric texts, cover the subject matter of this chapter more formally and in greater detail. Cox and Hinkley (1974) is a standard, if relatively difficult, treatment of most of the topics in this chapter. A compact summary appears in Zellner (1983). Wonnacott and Wonnacott (1990) present insightful treatments of many of these topics at a much lower level of mathematical sophistication; I particularly recommend this source if you found the simpler parts of this chapter too terse. A good, relatively accessible discussion of asymptotic distribution theory appears in Theil (1971, chap. 8). A general treatment of Wald, likelihood-ratio, and score tests can be found in Engle (1984). Finally, Lancaster (2004) presents an excellent and accessible extended introduction to Bayesian methods.

CHAPTER 4. PUTTING THE MATH TO WORK: LINEAR LEAST-SQUARES REGRESSION

As explained in the preface, this book aims to introduce mathematics useful for social statistics, and does not focus on statistical methods themselves. Nevertheless, I feel that it is helpful to convey some sense of how the math is employed to develop statistical methods. The purpose of the current chapter, therefore, is to illustrate this process by developing the statistical method of linear least-squares regression—a topic that I expect is at least somewhat familiar to the reader—and deriving some of its properties.

First, however, an important caveat: This chapter describes some of the mathematics of linear least squares, but it tells only part of the larger statistical story. Although mathematics plays an important role in applied statistics, applied statistics is not exclusively mathematical, and extends, for example, to methodological issues. Moreover, linear least-squares regression is in several respects the central method in applied statistics: It appears frequently in applications; extends readily to the general linear model, to generalized linear models, and beyond; and provides a computational basis for many statistical methods. Consequently, adequately explaining the role of linear least-squares regression in data analysis requires a much more extensive development of the topic than is possible in this chapter. This is the proper task of texts on applied regression analysis (such as my own: Fox, 2008).

The current chapter makes use of material in Chapter 1, on matrices and linear algebra, including matrix rank and the solution of linear simultaneous equations; in Chapter 2, on matrix differential calculus for optimization problems; and in Chapter 3, on probability, statistical distributions, properties of estimators, and maximum-likelihood estimation.

4.1 Least-Squares Fit

A linear regression equation can be written as

$$Y_i = A + B_1 x_{i1} + B_2 x_{i2} + \cdots + B_k x_{ik} + E_i \tag{4.1}$$

where Y_i is the value of a quantitative *response variable* (or "dependent variable") for the ith of n observations; $x_{i1}, x_{i2}, \ldots, x_{ik}$ are the values of k quantitative *explanatory variables* (or "independent variables") for

observation i; $A, B_1, B_2, \ldots, B_k$ are *regression coefficients*—the first of these, A, the regression *intercept* or *constant*, and the Bs, *partial slope coefficients*; and E_i is the regression *residual*, reflecting the departure of Y_i from the linear regression surface,

$$\widehat{Y}_i = A + B_1 x_{i1} + B_2 x_{i2} + \cdots + B_k x_{ik}$$

$\widehat{Y}_i$ is called the *fitted value* for observation i.

Notice that I use uppercase letters for Y_i and E_i; this usage reflects the fact that were we to draw a different sample of n observations, the values of the response variable, and, consequently, of the residuals would change; thus, Y_i and E_i are random variables. Similarly, because the values of the regression coefficients also change from sample to sample, they are also represented by uppercase letters. Conversely, I use lowercase letters for the explanatory variables to indicate that their values are fixed with respect to repeated sampling, a situation that typically occurs only in designed experiments, where the xs are under the direct control of the researcher and need not change if the study is replicated. Treating the xs as fixed produces simpler mathematics, and turns out to be nearly (but not quite) inconsequential; I will briefly consider random Xs in Section 4.6.

The least-squares regression coefficients are those values of A and the Bs that minimize the sum of squared residuals, considered as a function of the regression coefficients:

$$\begin{aligned} S(A, B_1, \ldots, B_k) &= \sum_{i=1}^{n} E_i^2 \\ &= \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2 \\ &= \sum_{i=1}^{n} (Y_i - A - B_1 x_{i1} - \cdots - B_k x_{ik})^2 \end{aligned}$$

Although we could continue with the linear regression equation in scalar form, it is advantageous to work instead with matrices. Let us therefore rewrite Equation 4.1 as

$$\underset{(n\times 1)}{\mathbf{y}} = \underset{(n\times k+1)}{\mathbf{X}} \underset{(k+1\times 1)}{\mathbf{b}} + \underset{(n\times 1)}{\mathbf{e}}$$

where $\mathbf{y} \equiv [Y_1, Y_2, \ldots, Y_n]'$ is a vector of observations on the response variable,

$$\mathbf{X} \equiv \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}$$

called the *model* (or *design*) *matrix*, contains the values of the explanatory variables, with an initial column of ones for the regression constant (called the *constant regressor*); $\mathbf{b} \equiv [A, B_1, \ldots, B_k]'$ contains the regression coefficients; and $\mathbf{e} \equiv [E_1, E_2, \ldots, E_n]'$ is a vector of residuals. Then the sum of squared residuals is

$$\begin{aligned} S(\mathbf{b}) &= \mathbf{e}'\mathbf{e} \\ &= (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) \\ &= \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\mathbf{b} - \mathbf{b}'\mathbf{X}'\mathbf{y} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b} \\ &= \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\mathbf{b} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b} \end{aligned} \tag{4.2}$$

The transition to the last line of Equation 4.2 is justified by the observation that $\mathbf{y}'\mathbf{X}\mathbf{b}$ is (1×1) and consequently is necessarily equal to its transpose, $\mathbf{b}'\mathbf{X}'\mathbf{y}$.

To minimize the sum-of-squares function $S(\mathbf{b})$, differentiate it with respect to the regression coefficients $\mathbf{b}$, a process that it facilitated by noting that Equation 4.2 consists of a constant (with respect to $\mathbf{b}$), a linear term in $\mathbf{b}$, and a quadratic form in $\mathbf{b}$; proceeding, we have

$$\frac{\partial S(\mathbf{b})}{\partial \mathbf{b}} = \mathbf{0} - 2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\mathbf{b}$$

Setting the vector partial derivative to $\mathbf{0}$ and rearranging produces the so-called *normal equations* for linear least-squares regression:

$$\underset{(k+1\times k+1)}{\mathbf{X}'\mathbf{X}} \quad \underset{(k+1\times 1)}{\mathbf{b}} = \underset{(k+1\times 1)}{\mathbf{X}'\mathbf{y}}$$

This is a system of $k + 1$ linear equations in the $k + 1$ unknown regression coefficients $\mathbf{b}$. The coefficient matrix for the system of equation,

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n & \sum x_{i1} & \sum x_{i2} & \cdots & \sum x_{ik} \\ \sum x_{i1} & \sum x_{ik}^2 & \sum x_{i1}x_{i2} & \cdots & \sum x_{i1}x_{ik} \\ \sum x_{i2} & \sum x_{i2}x_{i1} & \sum x_{i2}^2 & \cdots & \sum x_{i2}x_{ik} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum x_{ik} & \sum x_{ik}x_{i1} & \sum x_{ik}x_{i2} & \cdots & \sum x_{ik}^2 \end{bmatrix}$$

contains sum of squares and cross-products among the columns of the model matrix, while the right-hand-side vector, $\mathbf{X}'\mathbf{y} = \left[\sum Y_i, \sum x_{i1}Y_i, \sum x_{i2}Y_i, \ldots, \sum x_{ik}Y_i\right]'$ contains sums of cross-products between each column of the model matrix and the vector of responses. The sums of squares and products $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}'\mathbf{y}$ can be calculated directly from the data.

The $\mathbf{X}'\mathbf{X}$ matrix is of full rank, and hence nonsingular, if the model matrix $\mathbf{X}$ is of full column rank, $k + 1$—that is, if no explanatory variable is a perfect linear function of the others. Under these circumstances, the normal equations have the unique solution

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{4.3}$$

That this solution represents a minimum of $S(\mathbf{b})$ is supported by the fact that if $\mathbf{X}'\mathbf{X}$ is nonsingular it is also positive-definite.

4.2 A Statistical Model for Linear Regression

A common statistical model for linear regression is

$$Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

where, as before, Y_i represents the value of the response variable for the ith of n sample observations; also as before, $x_{i1}, x_{i2}, \ldots, x_{ik}$ are the values of the k explanatory variables; $\alpha, \beta_1, \beta_2, \ldots, \beta_k$ are population regression coefficients, to be estimated from the sample data; and ε_i is an *error variable* associated with observation i. A Greek letter is used for the error, even though it is a random variable, because the error is not directly observable. It is assumed that the errors are normally distributed with zero means and constant variance σ^2,

$$\varepsilon_i \sim N(0, \sigma^2)$$

and that errors from different observations are independent of one another.

Equivalently, in matrix form,

$$\underset{(n\times 1)}{\mathbf{y}} = \underset{(n\times k+1)}{\mathbf{X}}\underset{(k+1\times 1)}{\boldsymbol{\beta}} + \underset{(n\times 1)}{\boldsymbol{\varepsilon}} \tag{4.4}$$

where $\mathbf{y}$ is the response vector and $\mathbf{X}$ the model matrix, as in the preceding section; $\boldsymbol{\beta} \equiv [\alpha, \beta_1, \ldots, \beta_k]'$ is the vector of population regression

coefficients; and $\boldsymbol{\varepsilon} \equiv [\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n]'$ is the vector of errors. The error vector is multivariately normally distributed with a scalar covariance matrix, $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$. Notice that because they are independent, different errors are uncorrelated.[1]

The distribution of the response vector $\mathbf{y}$ follows from the distribution of $\boldsymbol{\varepsilon}$:

$$\begin{aligned}
\boldsymbol{\mu} \equiv E(\mathbf{y}) &= E(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\
&= \mathbf{X}\boldsymbol{\beta} + E(\boldsymbol{\varepsilon}) \\
&= \mathbf{X}\boldsymbol{\beta} \\
V(\mathbf{y}) &= E\left[(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})'\right] \\
&= E\left[(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} - \mathbf{X}\boldsymbol{\beta})(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} - \mathbf{X}\boldsymbol{\beta})'\right] \\
&= E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') \\
&= \sigma^2\mathbf{I}_n \\
\mathbf{y} &\sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)
\end{aligned}$$

Thus, for example, the assumption that $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ implies that $E(\mathbf{y})$ is a linear function of $\mathbf{X}$.

4.3 The Least-Squares Coefficients as Estimators

The least-squares regression coefficients $\mathbf{b}$ from Equation 4.3 (page 155) may be used to estimate the coefficients of the linear regression model of Equation 4.4. Because $\mathbf{b}$ results from a linear transformation of the response vector $\mathbf{y}$, the properties of the least-squares estimator are easily established:

$$\mathbf{b} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} = \mathbf{M}\mathbf{y}$$

where the transformation matrix $\mathbf{M} \equiv \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$. Because the model matrix $\mathbf{X}$ is fixed with respect to repeated sampling, so is $\mathbf{M}$. Then,

$$E(\mathbf{b}) = \mathbf{M}E(\mathbf{y}) = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

[1] As a general matter, independence implies uncorrelation, but uncorrelated random variables are not necessarily independent (see Section 3.1.2). In the multivariate-normal distribution, however, independence and uncorrelation coincide.

demonstrating that $\mathbf{b}$ is an unbiased estimator of $\boldsymbol{\beta}$. Note that this conclusion depends only on the assumption that $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ (i.e., the assumption of linearity).

The covariance matrix of $\mathbf{b}$ is also simply derived from the assumptions of constant error variance and uncorrelated errors [i.e., that $V(\mathbf{y}) = \sigma^2\mathbf{I}_n$]:

$$\begin{aligned} V(\mathbf{b}) &= \mathbf{M}V(\mathbf{y})\mathbf{M}' \\ &= [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\sigma^2\mathbf{I}_n[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']' \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

Finally, from the assumption of normally distributed errors,

$$\mathbf{b} \sim N_{k+1}\left[\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right] \tag{4.5}$$

It can be shown that the least-squares estimator $\mathbf{b}$ is not only an unbiased estimator of $\boldsymbol{\beta}$, but, under the assumptions of linearity, constant error variance, and independence, it is the minimum-variance unbiased estimator that is a linear function of the data. This result, called the *Gauss-Markov theorem*, is often taken as justification for least-squares estimation but it does not lend strong support to the least-squares estimator: When the error distribution is non-normal, other unbiased estimators that are *nonlinear* functions of the data (so-called *robust-regression* estimators) can be much more efficient than the least-squares estimator. When the errors are normally distributed, however, the least-squares estimator is maximally efficient among *all* unbiased estimators—a much more compelling result.[2] The Gauss-Markov theorem is named after the great 18th–19th century German mathematician Carl Friedrich Gauss, and the 19th–20th century Russian probabilist Andrey Andreevich Markov.

4.4 Statistical Inference for the Regression Model

Statistical inference for the population regression coefficients $\boldsymbol{\beta}$, beyond point estimation, is complicated by the fact that we typically do not know the error variance σ^2 and therefore cannot directly apply Equation 4.5 for

[2]For proofs of the results mentioned in this paragraph, see, for example, Rao (1973).

the distribution of the least-squares estimator $\mathbf{b}$. We must instead estimate σ^2 along with $\boldsymbol{\beta}$.

An unbiased estimator of σ^2 is given by

$$S^2 = \frac{\sum_{i=1}^{n} E_i^2}{n-k-1} = \frac{\mathbf{e}'\mathbf{e}}{n-k-1}$$

where $n-k-1$ are the degrees of freedom for error (having "lost" $k+1$ degrees of freedom as a consequence of estimating the $k+1$ elements of $\boldsymbol{\beta}$). Then the estimated covariance matrix for the least-squares coefficients is

$$\widehat{V}(\mathbf{b}) = S^2(\mathbf{X}'\mathbf{X})^{-1}$$

and the square-roots of the diagonal entries of $\mathbf{b}$ are the standard errors of the regression coefficients, $\mathrm{SE}(A)$, $\mathrm{SE}(B_1), \cdots, \mathrm{SE}(B_k)$.

Inference for individual regression coefficients is based on the t-distribution. For example, to test the hypothesis H_0: $\beta_j = \beta_j^{(0)}$ that a slope coefficient is equal to the particular value $\beta_j^{(0)}$ (typically 0), we compute the test statistic

$$t_0 = \frac{B_j - \beta_j^{(0)}}{\mathrm{SE}(B_j)}$$

which is distributed as t_{n-k-1} under H_0. Similarly, to construct a 95% confidence interval for β_j, we take

$$\beta_j = B_j \pm t_{n-k-1,.025}\mathrm{SE}(B_j)$$

where $t_{n-k-1,.025}$ is the critical value of t with $n-k-1$ degrees of freedom and a probability of .025 to the right.

More generally, we can test the linear hypothesis

$$H_0\colon \underset{(q\times k+1)}{\mathbf{L}}\ \underset{(k+1\times 1)}{\boldsymbol{\beta}} = \underset{(q\times 1)}{\mathbf{c}}$$

where $\mathbf{L}$ and $\mathbf{c}$ contain prespecified constants, and the *hypothesis matrix* $\mathbf{L}$ is of full row rank $q \leq k+1$. The resulting F-statistic,

$$F_0 = \frac{(\mathbf{Lb}-\mathbf{c})'\left[\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}'\right]^{-1}(\mathbf{Lb}-\mathbf{c})}{qS^2}$$

follows an F-distribution with q and $n-k-1$ degrees of freedom if H_0 is true.

Suppose, for example, that we wish to test the "omnibus" null hypothesis H_0: $\beta_1 = \beta_2 = 0$ in a regression model with two explanatory variables; we can take

$$\mathbf{L} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

and $\mathbf{c} = [0, 0]'$. To test the hypothesis that the regression coefficients are equal, H_0: $\beta_1 = \beta_2$, which is equivalent to H_0: $\beta_1 - \beta_2 = 0$, we can take $\mathbf{L} = [0, 1, -1]$ and $\mathbf{c} = [0]$.[3]

As shown in the next section, under the assumptions of the regression model, the least-squares estimators of the regression coefficients are maximum-likelihood estimators. Consequently, if the sample sample size is sufficiently large, we can employ the delta method (Section 3.6.5) to derive the standard error of a *nonlinear* function of the regression coefficients.

Consider, for example, the quadratic regression model

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon \tag{4.6}$$

(This model can be fit by linear-least-squares regression of Y on x and x^2 because it is linear in the parameters β_0, β_1, and β_2.) Suppose that we are interested in determining the x-value at which the regression equation reaches its maximum or minimum.[4] Taking the expectation of both sides of Equation 4.6 and differentiating with respect to x (see Section 2.4), we get

$$\frac{dE(Y)}{dx} = \beta_1 + 2\beta_2 x$$

Setting the derivative to 0 and solving for x produces the x-value at which the function reseaches a minimum (if β_2 is positive) or a maximum (if β_2 is negative),

$$x = -\frac{\beta_1}{2\beta_2}$$

which is a nonlinear function of the regression coefficients β_1 and β_2.

---

[3]For this hypothesis to be sensible, the two explanatory variables x_1 and x_2, would have to be measured in the same units.

[4]The application of the delta method to this problem is suggested by Weisberg (2005, sect. 6.1.2).

To apply the delta method, we need the partial derivatives of $\gamma = f(\beta_1, \beta_2) \equiv -\beta_1/(2\beta_2)$ with respect to the regression coefficients:

$$\frac{\partial \gamma}{\partial \beta_1} = -\frac{1}{2\beta_2}$$
$$\frac{\partial \gamma}{\partial \beta_2} = \frac{\beta_1}{2\beta_2^2}$$

Now suppose that we compute least-squares estimates B_1 of β_1 and B_2 of β_2, along with the estimated variances of the coefficients, $\widehat{V}(B_1)$ and $\widehat{V}(B_2)$, and their covariance, $\widehat{C}(B_1, B_2)$. The maximum-likelihood estimate of γ is $\widehat{\gamma} = -B_1/(2B_2)$, and the delta-method variance of $\widehat{\gamma}$ is

$$\widehat{V}(\widehat{\gamma}) = \widehat{V}(B_1)\left(-\frac{1}{2B_2}\right)^2 + \widehat{V}(B_2)\left(\frac{B_1}{2B_2^2}\right)^2 + 2\widehat{C}(B_1, B_2)\left(-\frac{1}{2B_2}\right)\left(\frac{B_1}{2B_2^2}\right)$$

Thus, a 95% confidence interval for γ is given by $\widehat{\gamma} \pm 1.96\sqrt{\widehat{V}(\widehat{\gamma})}$.

4.5 Maximum-Likelihood Estimation of the Regression Model

As I have explained, under the assumptions of the linear model, $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$. Thus, for the ith observation, $Y_i \sim N(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2)$, where $\mathbf{x}_i'$ is the ith row of the model matrix $\mathbf{X}$. In equation form, the probability density for observation i is

$$p(y_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{2\sigma^2}\right]$$

Because the n observations are independent, their joint probability density is the product of their marginal densities:

$$p(\mathbf{y}) = \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n} \exp\left[-\frac{\sum (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{2\sigma^2}\right] \tag{4.7}$$
$$= \frac{1}{\left(2\pi\sigma^2\right)^{n/2}} \exp\left[-\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2}\right]$$

Although this equation also follows directly from the multivariate-normal distribution of $\mathbf{y}$, the development from $p(y_i)$ to $p(\mathbf{y})$ will prove helpful when we consider random regressors (in the next section).

From Equation 4.7, the log likelihood is

$$\log_e L(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log_e 2\pi - \frac{n}{2}\log_e \sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{4.8}$$

To maximize the likelihood, we require the partial derivatives of Equation 4.8 with respect to the parameters $\boldsymbol{\beta}$ and σ^2. Differentiation is simplified when we notice that $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ is the sum of squared errors:

$$\frac{\partial \log_e L(\boldsymbol{\beta}, \sigma^2)}{\partial \boldsymbol{\beta}} = -\frac{1}{2\sigma^2}(2\mathbf{X}'\mathbf{X}\boldsymbol{\beta} - 2\mathbf{X}'\mathbf{y})$$

$$\frac{\partial \log_e L(\boldsymbol{\beta}, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2}\left(\frac{1}{\sigma^2}\right) + \frac{1}{2\sigma^4}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Setting these partial derivatives to 0 and solving for the maximum-likelihood estimators $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}^2$ produces

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

$$\widehat{\sigma}^2 = \frac{(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})}{n} = \frac{\mathbf{e}'\mathbf{e}}{n}$$

The maximum-likelihood estimator $\widehat{\boldsymbol{\beta}}$ is therefore the same as the least-squares estimator $\mathbf{b}$. In fact, this identity is clear directly from Equation 4.7, without formal maximization of the likelihood: The likelihood is large when the negative exponent is small, and the numerator of the exponent contains the sum of squared errors; minimizing the sum of squared residuals, therefore, maximizes the likelihood.

The maximum-likelihood estimator $\widehat{\sigma}^2$ of the error variance is biased; consequently, we prefer the similar, unbiased estimator $S^2 = \mathbf{e}'\mathbf{e}/(n - k - 1)$, described previously. As n increases, however, the bias of $\widehat{\sigma}^2$ shrinks toward 0: As a maximum-likelihood estimator, $\widehat{\sigma}^2$ is consistent.

4.6 Random Xs

The theory of linear regression analysis developed in this chapter has proceeded from the premise that the model matrix $\mathbf{X}$ is *fixed*. If we repeat a study, we expect the response-variable observations $\mathbf{y}$ to change, but if $\mathbf{X}$ is fixed, then the explanatory-variable values are constant across replications of the study. This situation is realistically descriptive of an experiment, where the

explanatory variables are manipulated by the researcher. Most research in the social sciences, however, is observational rather than experimental; and in an observational study (survey research, for example), we would typically obtain different explanatory-variable values on replication of the study. In observational research, therefore, $\mathbf{X}$ is *random* rather than fixed.

It is remarkable that the statistical theory of linear regression applies even when $\mathbf{X}$ is random, as long as certain assumptions are met. For fixed explanatory variables, the assumptions underlying the model take the form $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$. That is, the distribution of the error is the same for all observed combinations of explanatory-variable values represented by the distinct rows of the model matrix. When $\mathbf{X}$ is random, we need to assume that this property holds for *all possible* combinations of explanatory-variable values in the population that is sampled: That is, $\mathbf{X}$ and $\boldsymbol{\varepsilon}$ are assumed to be independent, and thus the *conditional* distribution of the error for a sample of explanatory variable values $\boldsymbol{\varepsilon}|\mathbf{X}_0$ is $N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$, regardless of the *particular* sample $\mathbf{X}_0 = \{x_{ij}\}$ that is chosen.

Because $\mathbf{X}$ is random, it has some (multivariate) probability distribution. It is not necessary to make assumptions about this distribution, however, beyond (1) requiring that $\mathbf{X}$ is measured without error, and that $\mathbf{X}$ and $\boldsymbol{\varepsilon}$ are independent (as just explained); (2) assuming that the distribution of $\mathbf{X}$ does not depend on the parameters $\boldsymbol{\beta}$ and σ^2 of the regression model; and (3) stipulating that the covariance matrix of the Xs is nonsingular (i.e., that no X is invariant or a perfect linear function of the others in the population). In particular, we need *not* assume that the *regressors* (as opposed to the *errors*) are normally distributed. This is fortunate, for many Xs are highly non-normal—dummy regressors and polynomial regressors come immediately to mind, not to mention many quantitative explanatory variables.[5]

It would be unnecessarily tedious to recapitulate the entire argument of this chapter, but I will show that some key results hold under the new assumptions

[5]To say that the previous results hold with random Xs under these new assumptions is not to assert that the new assumptions are necessarily unproblematic. Most explanatory variables are measured with error, and to assume otherwise can, in certain circumstances, seriously bias the estimated regression coefficients. Similarly, under certain (causal) interpretations of regression equations, the assumption that the errors are independent of the explanatory variables is tantamount to assuming that the aggregated omitted determinants of Y are unrelated to the determinants of Y that are included in the model. Finally, the assumptions of linearity, constant error variance, and normality are also potentially problematic. Dealing satisfactorily with these issues is the difference between regression analysis as a mathematical abstraction and as a practical tool for data analysis.

when the explanatory variables are random. The other results of the chapter can be established for random Xs in a similar manner.

For a particular sample of X-values, $\mathbf{X}_0$, the conditional distribution of $\mathbf{y}$ is

$$\begin{aligned} E(\mathbf{y}|\mathbf{X}_0) &= E\left[(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})|\mathbf{X}_0\right] \\ &= \mathbf{X}_0\boldsymbol{\beta} + E(\boldsymbol{\varepsilon}|\mathbf{X}_0) \\ &= \mathbf{X}_0\boldsymbol{\beta} \end{aligned}$$

Consequently, the conditional expectation of the least-squares estimator is

$$\begin{aligned} E(\mathbf{b}|\mathbf{X}_0) &= E\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}|\mathbf{X}_0\right] \\ &= (\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{X}_0'E(\mathbf{y}|\mathbf{X}_0) \\ &= (\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{X}_0'\mathbf{X}_0\boldsymbol{\beta} \\ &= \boldsymbol{\beta} \end{aligned}$$

Because we can repeat this argument for *any* value of $\mathbf{X}$, the least-squares estimator $\mathbf{b}$ is conditionally unbiased for any and every such value; it is therefore *unconditionally* unbiased as well, $E(\mathbf{b}) = \boldsymbol{\beta}$.

Suppose now that we use the procedures of Section 4.4 to perform statistical inference for $\boldsymbol{\beta}$. For concreteness, imagine that we calculate a p-value for the omnibus null hypothesis H_0: $\beta_1 = \cdots = \beta_k = 0$. Because $\boldsymbol{\varepsilon}|\mathbf{X}_0 \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$, as was required when we treated $\mathbf{X}$ as fixed, the p-value obtained is correct for $\mathbf{X} = \mathbf{X}_0$ (i.e., for the sample at hand). There is, however, nothing special about a particular $\mathbf{X}_0$: The error vector $\boldsymbol{\varepsilon}$ is independent of $\mathbf{X}$, and so the distribution of $\boldsymbol{\varepsilon}$ is $N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$ for any and every value of $\mathbf{X}$. The p-value, therefore, is *unconditionally* valid.

Finally, I will show that the maximum-likelihood estimators of $\boldsymbol{\beta}$ and σ^2 are unchanged when $\mathbf{X}$ is random, as long as the new assumptions hold: When $\mathbf{X}$ is random, sampled observations consist not just of response-variable values $(Y_1, \ldots, Y_n)$ but also of explanatory-variable values $(\mathbf{x}_1', \ldots, \mathbf{x}_n')$. The observations themselves are denoted $[Y_1, \mathbf{x}_1'], \ldots, [Y_n, \mathbf{x}_n']$. Because these observations are sampled independently, their joint probability density is the product of their marginal densities:

$$p(\mathbf{y}, \mathbf{X}) \equiv p\left([y_1, \mathbf{x}_1'], \ldots, [y_n, \mathbf{x}_n']\right) = p(y_1, \mathbf{x}_1') \times \cdots \times p(y_n, \mathbf{x}_n')$$

Now, the probability density $p(y_i, \mathbf{x}_i')$ for observation i can be written as $p(y_i|\mathbf{x}_i')\,p(\mathbf{x}_i')$. According to the linear model, the conditional distribution of Y_i given $\mathbf{x}_i'$ is normal:

$$p(y_i|\mathbf{x}_i') = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{2\sigma^2}\right]$$

Thus, the joint probability density for all observations becomes

$$\begin{aligned} p(\mathbf{y}, \mathbf{X}) &= \prod_{i=1}^{n} p(\mathbf{x}_i') \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{2\sigma^2}\right] \\ &= \left[\prod_{i=1}^{n} p(\mathbf{x}_i')\right] \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2}\right] \\ &= p(\mathbf{X})\,p(\mathbf{y}|\mathbf{X}) \end{aligned}$$

As long as $p(\mathbf{X})$ does not depend on the parameters $\boldsymbol{\beta}$ and σ^2, we can ignore the joint density of the Xs in maximizing $p(\mathbf{y}, \mathbf{X})$ with respect to the parameters. Consequently, the maximum-likelihood estimator of $\boldsymbol{\beta}$ is the least-squares estimator, as was the case for fixed $\mathbf{X}$.

REFERENCES

Aldrich, J. (1997). R. A. Fisher and the making of maximum-likelihood 1912–1922. *Statistical Science, 12*, 162–176.

Binmore, K., & Davies, J. (2001). *Calculus: Concepts and methods*. Cambridge, UK: Cambridge University Press.

Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics*. London: Chapman & Hall.

Davis, P. J. (1973). *The mathematics of matrices: A first book of matrix theory and linear algebra* (2nd ed.). Lexington, MA: Xerox College.

Engle, R. F. (1984). Wald, likelihood ratio, and Lagrange multiplier tests in econometrics. In Z. Griliches & M. D. Intriligator (Eds.), *Handbook of econometrics* (Vol. 2, pp. 775–879). Amsterdam: North-Holland.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London A, 222*, 309–368.

Fox, J. (2008). *Applied regression analysis and generalized linear models* (2nd ed.). Thousand Oaks, CA: Sage.

Graybill, F. A. (1983). *Introduction to matrices with applications in statistics* (2nd ed.). Belmont, CA: Wadsworth.

Green, P. E., & Carroll, J. D. (1976). *Mathematical tools for applied multivariate analysis*. New York: Academic Press.

Healy, M. J. R. (1986). *Matrices for statistics*. Oxford, UK: Clarendon Press.

Johnston, J. (1972). *Econometric methods* (2nd ed.). New York: McGraw-Hill.

Kennedy, W. J., Jr., & Gentle, J. E. (1980). *Statistical computing*. New York: Dekker.

Lancaster, T. (2004). *An introduction to modern Bayesian econometrics*. Oxford, UK: Blackwell.

McCallum, B. T. (1973). A note concerning asymptotic covariance expressions. *Econometrica, 41*, 581–583.

Monahan, J. F. (2001). *Numerical methods of statistics*. Cambridge, UK: Cambridge University Press.

Namboodiri, K. (1984). *Matrix algebra: An introduction*. Beverly Hills, CA: Sage.

Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York: Wiley.

Rao, C. R., & Mitra, S. K. (1971). *Generalized inverse of matrices and its applications*. New York: Wiley.

Searle, S. R. (1982). *Matrix algebra useful for statistics*. New York: Wiley.

Theil, H. (1971). *Principles of econometrics*. New York: Wiley.

Thompson, S. P., & Gardner, M. (1998). *Calculus made easy*. New York: St. Martin's Press.

Weisberg, S. (2005). *Applied linear regression* (3rd ed.). New York: Wiley.

Wonnacott, T. H., & Wonnacott, R. J. (1990). *Introductory statistics* (5th ed.). New York: Wiley.

Zellner, A. (1983). Statistical theory and econometrics. In Z. Griliches & M. D. Intriligator (Eds.), *Handbook of econometrics* (Vol. 1, pp. 67–178). Amsterdam: North-Holland.

INDEX