Appendices to *Applied Regression Analysis, Generalized Linear Models, and Related Methods, Second Edition*

John Fox[1]
Department of Sociology
McMaster University
jfox@mcmaster.ca

7 January 2008

ii

Contents

Preface to the Appendices

These appendices are meant to accompany my text on *Applied Regression, Generalized Linear Models, and Related Methods, Second Edition* (Sage, 2007). Appendix A on *Notation*, which appears in the printed text, is reproduced in slightly expanded form here for convenience. The other appendices are available only in this document. Appendices B (on *Matrices, Linear Algebra, and Vector Geometry*) and C (on *Calculus*) are starred not because they are terribly difficult but because they are required only for starred portions of the main text. Parts of Appendix D (on *Probabilty and Estimation*) are left un-starred because they are helpful for some un-starred material in the main text.

Individuals who do not have a copy of my *Applied Regression* text are welcome to read these appendices if they find them useful, but please do not ask me questions about them. Of course, I would be grateful to learn of any errors.

Appendix A

Notation

Specific notation is introduced at various points in the appendices and chapters. 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 $[a, b, x_i, x_1^*]$.

- Observable scalar random variables are represented by uppercase italic letters $[X, Y_i, B_0']$ or, if the names contain more than one character, by roman letters, the first of which is uppercase [RegSS, RSS$_0$]. Where it is necessary to make the distinction, *specific values* of random variables are represented as constants $[x, y_i, b_0']$.

- Scalar parameters are represented by lowercase Greek letters $[\alpha, \beta, \beta_j^*, \gamma_2]$. (See the Greek alphabet in Table A.1.) Their estimators are generally denoted by "corresponding" italic characters $[A, B, B_j^*, C_2]$, or by Greek letters with diacritics $[\widehat{\alpha}, \widehat{\beta}]$.

- Unobservable scalar random variables are also represented by lowercase Greek letters $[\varepsilon_i]$.

- Vectors and matrices are represented by boldface characters—lowercase for vectors $[\mathbf{x}_1, \boldsymbol{\beta}]$, uppercase for matrices $[\mathbf{X}, \boldsymbol{\Sigma}_{12}]$. Roman letters are used for constants and observable random variables $[\mathbf{y}, \mathbf{x}_1, \mathbf{X}]$. Greek letters are used for parameters and unobservable random variables $[\boldsymbol{\beta}, \boldsymbol{\Sigma}_{12}, \boldsymbol{\varepsilon}]$. 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 1's is represented by a boldface 1, possibly subscripted with its number of elements $[\mathbf{1}_n]$. Vectors are column vectors, unless they are explicitly transposed [column: $\mathbf{x}$; row: $\mathbf{x}'$].

- Diacritics and symbols such as * (asterisk) and $'$ (prime) are used freely as modifiers to denote alternative forms $[\mathbf{X}^*, \beta', \widetilde{\varepsilon}]$.

- The symbol $\equiv$ can be read as "is defined by," or "is equal to by definition" $[\overline{X} \equiv (\sum X_i)/n]$.

1

Table A.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

- The symbol $\approx$ means "is approximately equal to" $[1/3 \approx 0.333]$.

- The symbol $\propto$ means "is proportional to" $[p(\alpha|D) \propto L(\alpha)p(\alpha)]$.

- The symbol $\sim$ means "is distributed as" $[\varepsilon_i \sim N(0, \sigma_\varepsilon^2)]$.

- The symbol $\in$ denotes membership in a set $[1 \in \{1, 2, 3\}]$.

- 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 $[C(X, Y), C(\mathbf{x}_i, \boldsymbol{\varepsilon})]$.

- 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 $[\text{plim } b = \beta]$.

- Standard mathematical functions are shown in lowercase $[\cos W, \text{trace}(\mathbf{A})]$. 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 $\#$ indicates a count $[\#_{i=1}^n(T_b^* \geq T)]$.

- 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."]

Matrices, Linear Algebra, and Vector Geometry*

Matrices provide a natural notation for linear models and, indeed, much of statistics; the algebra of linear 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 appendix is to present basic concepts and results concerning matrices, linear algebra, and vector geometry. The focus is on topics that are employed in the main body of the book, 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 appendix).

The first section of the appendix develops elementary matrix algebra. Sections B.2 and B.3 introduce vector geometry and vector spaces. Section B.4 discusses the related topics of matrix rank and the solution of linear simultaneous equations. Sections B.5 and B.6 deal with eigenvalues, eigenvectors, quadratic forms, and positive-definite matrices.

B.1 Matrices

B.1.1 Introducing the Actors: Basic Definitions

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

$$
\underset{(4\times3)}{\mathbf{X}} = \begin{bmatrix} 1 & -2 & 3 \\ 4 & -5 & -6 \\ 7 & 8 & 9 \\ 0 & 0 & 10 \end{bmatrix} \tag{B.1}
$$

or, more generally,

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

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 B.1 and B.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 B.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 often 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 B.1 and B.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} (that is, $a_{11}, a_{22}, \ldots, a_{nn}$) 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 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:

$$\mathop{\mathbf{U}}_{(n\times n)} = \left[\begin{array}{cccc} u_{11} & u_{12} & \cdots & u_{1n} \\ 0 & u_{22} & \cdots & u_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & u_{nn} \end{array} \right]$$

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

$$\mathop{\mathbf{L}}_{(n\times n)} = \left[\begin{array}{cccc} l_{11} & 0 & \cdots & 0 \\ l_{21} & l_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & \cdots & l_{nn} \end{array} \right]$$

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

$$\mathop{\mathbf{D}}_{(n\times n)} = \left[\begin{array}{cccc} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_n \end{array} \right]$$

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:

$$\mathop{\mathbf{I}}_{(n\times n)} = \left[\begin{array}{cccc} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{array} \right]$$

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

Two other special matrices are the family of *zero matrices* $\mathbf{0}$, all of whose entries are zero, and the unit vectors $\mathbf{1}$, all of whose entries are one. I write $\mathbf{1}_n$ for the unit vector with n entries; for example $\mathbf{1}_4 = (1, 1, 1, 1)'$. Although the identity matrices, the zero matrices, and the unit vectors are *families* of matrices, it is often convenient to refer to these matrices in the singular, for example, to *the* identity matrix.

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

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

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, $\underset{(m \times n+p)}{\mathbf{C}} =$

$$\begin{bmatrix} \underset{(m \times n)}{\mathbf{C}_1} , & \underset{(m \times p)}{\mathbf{C}_2} \end{bmatrix}.$$

B.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 B.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

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

and

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

we have

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

$$\underset{(2 \times 3)}{\mathbf{D}} = \mathbf{A} - \mathbf{B} = \begin{bmatrix} 6 & 1 & 1 \\ 1 & 5 & 10 \end{bmatrix}$$

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

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:

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

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

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

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

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

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

For example,

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

Two matrices A and B are *conformable for multiplication* in the order given (i.e., AB) if the number of *columns* of the left-hand factor (A) is equal to the number of *rows* of the right-hand factor (B). Thus A and B are conformable for multiplication if A is of order $(m \times n)$ and 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 $C = AB$ be the matrix product; and let a'_i represent the ith *row* of A and b_j represent the jth *column* of B. Then C is a matrix of order $(m \times p)$ in which

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

---

[1] 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.

Some examples:

$$
\begin{bmatrix}
 & \Longrightarrow & \\
1 & 2 & 3 \\
4 & 5 & 6
\end{bmatrix}_{(2\times3)}
\begin{bmatrix}
 & 1 & 0 & 0 \\
\Downarrow & 0 & 1 & 0 \\
 & 0 & 0 & 1
\end{bmatrix}_{(3\times3)}
$$

$$
=
\begin{bmatrix}
1(1) + 2(0) + 3(0), & 1(0) + 2(1) + 3(0), & 1(0) + 2(0) + 3(1) \\
4(1) + 5(0) + 6(0), & 4(0) + 5(1) + 6(0), & 4(0) + 5(0) + 6(1)
\end{bmatrix}_{(2\times3)}
$$

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

$$
[\beta_0, \beta_1, \beta_2, \beta_3]_{(1\times4)}
\begin{bmatrix}
1 \\
x_1 \\
x_2 \\
x_3
\end{bmatrix}_{(4\times1)}
= [\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3]_{(1\times1)}
$$

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

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

$$
\begin{bmatrix}
2 & 0 \\
0 & 3
\end{bmatrix}
\begin{bmatrix}
\frac{1}{2} & 0 \\
0 & \frac{1}{3}
\end{bmatrix}
=
\begin{bmatrix}
1 & 0 \\
0 & 1
\end{bmatrix}
\tag{B.4}
$$

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

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

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

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 B.3). If it is the case that $\mathbf{AB} = \mathbf{BA}$ (as in Equation B.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: $c\mathbf{AB} = \mathbf{A}c\mathbf{B} = \mathbf{AB}c$.

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

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

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

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

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:

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

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.[2] If $\mathbf{A}^2 = \mathbf{A}$, then $\mathbf{A}$ is said to be *idempotent*.

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

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

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

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

and

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

then

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

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:

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

---

[2] Of course, even the scalar square-root is unique only up to a change in sign.

Writing these equations as a matrix equation,

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

$$\underset{(2\times2)}{\mathbf{A}}\ \underset{(2\times1)}{\mathbf{x}} = \underset{(2\times1)}{\mathbf{b}}$$

The formulation and solution of systems of linear simultaneous equations is taken up in Section B.4.

B.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[3] $\mathbf{A}$ 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*.[4] If the inverse of a matrix exists, then it is unique; moreover, if for a square matrix $\mathbf{A}$, $\mathbf{A}\mathbf{B} = \mathbf{I}$, then necessarily $\mathbf{B}\mathbf{A} = \mathbf{I}$, 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:

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

---

[3]It is possible to define various sorts of *generalized inverses* for rectangular matrices and for square matrices that do not have conventional inverses. Although generalized inverses have statistical applications, I do not use them in the text. See, for example, Rao and Mitra (1971).

[4]When mathematicians first encountered nonzero matrices without inverses, they found this result remarkable or "singular."

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 **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 **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*.[5] 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 appendix). To illustrate the method of elimination, I will employ the matrix

$$\begin{bmatrix} 2 & -2 & 0 \\ 1 & -1 & 1 \\ 4 & 4 & -4 \end{bmatrix} \tag{B.5}$$

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:

Divide row 1 by 2,

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

Subtract the "new" row 1 from row 2,

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

---

[5] After the great German mathematician, Carl Friedrich Gauss (1777–1855).

Subtract $4 \times$ row 1 from row 3,

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

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

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

Divide row 2 by 8,

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

Add row 2 to row 1,

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

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

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

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:

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

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

$$\mathbf{E}_{III} \equiv \left[\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{array}\right]$$

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

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

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 $\mathbf{E}\left[\mathbf{A}, \mathbf{I}_n\right] = \left[\mathbf{I}_n, \mathbf{B}\right]$; that is, $\mathbf{E}\mathbf{A} = \mathbf{I}_n$ (implying that $\mathbf{E} = \mathbf{A}^{-1}$), and $\mathbf{E}\mathbf{I}_n = \mathbf{B}$. 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:

$$\mathbf{I}^{-1} = \mathbf{I}$$
$$(\mathbf{A}^{-1})^{-1} = \mathbf{A}$$
$$(\mathbf{A}')^{-1} = (\mathbf{A}^{-1})'$$
$$(\mathbf{A}\mathbf{B})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$$
$$(c\mathbf{A})^{-1} = c^{-1}\mathbf{A}^{-1}$$

(where $\mathbf{A}$ and $\mathbf{B}$ are 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.

B.1.4 Determinants

Each square matrix $\mathbf{A}$ is associated with a scalar called its *determinant*, written $\det \mathbf{A}$. 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

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

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 B.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 B.5 (on page 13), then, the determinant is $-(2)(8)(1) = -16$. 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.

B.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

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

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

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

because

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

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

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

B.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

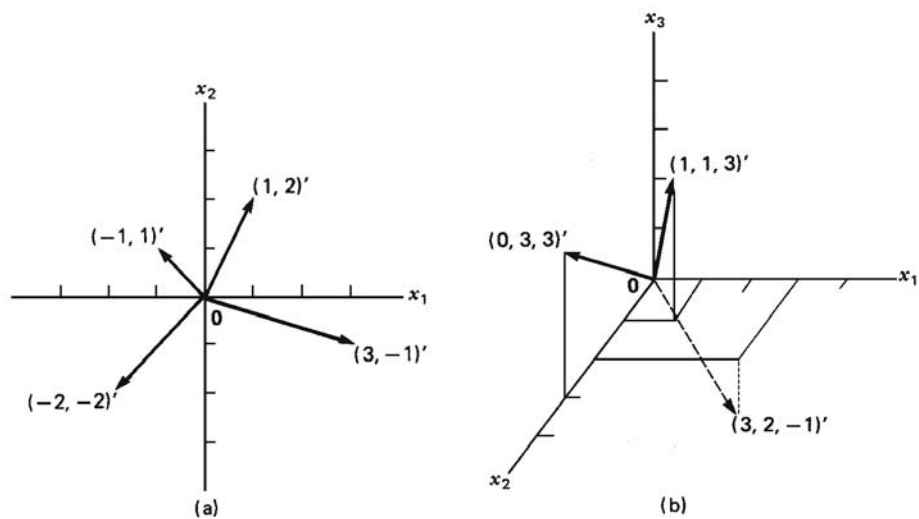Figure B.1: Examples of geometric vectors in (a) two-dimensional and (b) three-dimensional space. Each vector is a directed line segment from the origin (**0**) to the point whose coordinates are given by the entries of the vector.
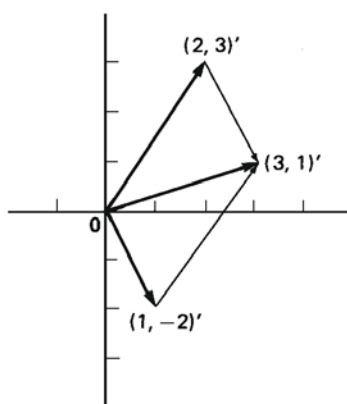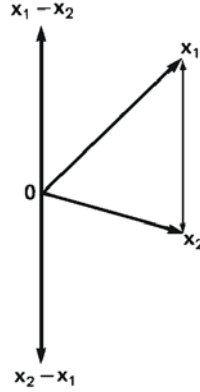

Figure B.2: Vectors are added by placing the "tail" of one on the tip of the other and completing the parallelogram. The sum is the diagonal of the parallelogram starting at the origin.

Figure B.3: Vector differences $\mathbf{x}_1 - \mathbf{x}_2$ and $\mathbf{x}_2 - \mathbf{x}_1$.

vector. Some examples of geometric vectors in two- and three-dimensional space are shown in Figure B.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" 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 B.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 B.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 B.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 B.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 B.3).

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
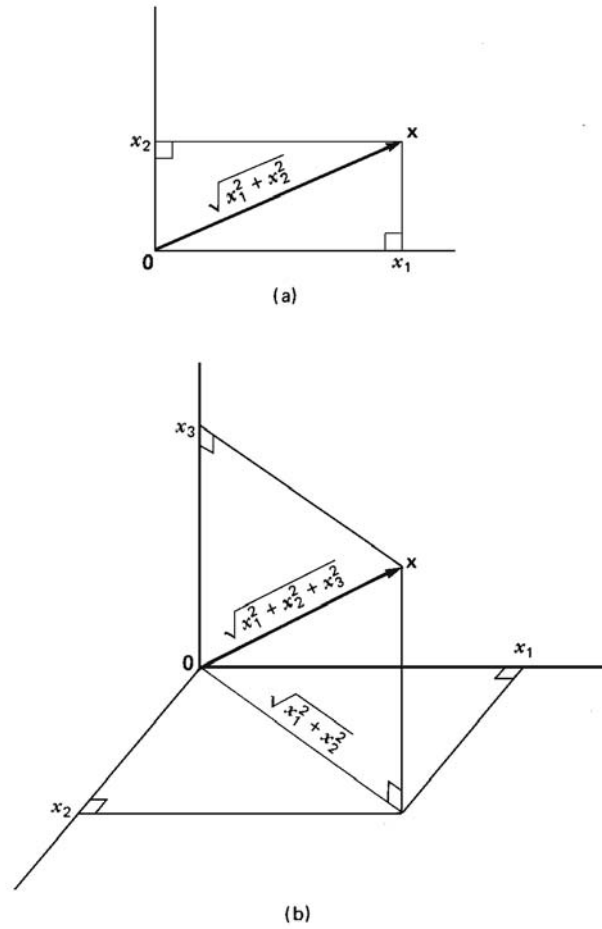
Figure B.4: The length of a vector is the square root of its sum of squared coordinates, $||\mathbf{x}|| = \sqrt{\sum_{i=1}^{n} x_i^2}$. This result is illustrated in (a) two and (b) three dimensions.
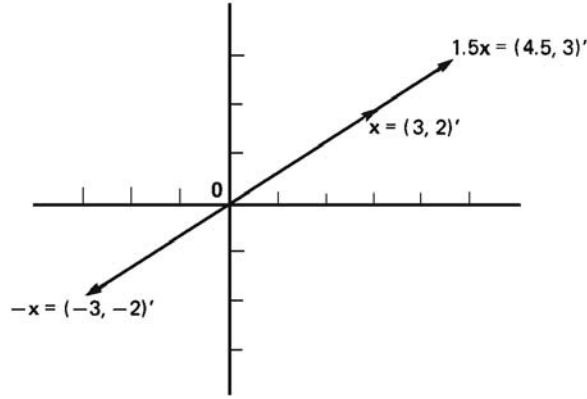
Figure B.5: Product $a\mathbf{x}$ of a scalar and a vector, illustrated in two dimensions. The vector $a\mathbf{x}$ is collinear with $\mathbf{x}$; it is in the same direction as $\mathbf{x}$ if $a > 0$, and in the opposite direction from $\mathbf{x}$ if $a < 0$.

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 B.5.

B.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.

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:[6]

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

---

[6]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.

The set of vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$ is said to *span* the subspace that it generates.

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:

$$\mathbf{x}_j = a_1\mathbf{x}_1 + \cdots + a_{j-1}\mathbf{x}_{j-1} + a_{j+1}\mathbf{x}_{j+1} + \cdots + a_k\mathbf{x}_k \tag{B.6}$$

(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

$$b_1\mathbf{x}_1 + b_2\mathbf{x}_2 + \cdots + b_k\mathbf{x}_k = \underset{(n \times 1)}{\mathbf{0}} \tag{B.7}$$

Equation B.6 or B.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 B.6. Figure B.6(a) shows the one-dimensional subspace (i.e., the line) generated by a single nonzero vector $\mathbf{x}$; Figure B.6(b) shows the one-dimensional subspace generated by two collinear vectors $\mathbf{x}_1$ and $\mathbf{x}_2$; Figure B.6(c) shows the two-dimensional subspace (the plane) generated by two linearly independent vectors $\mathbf{x}_1$ and $\mathbf{x}_2$; and Figure B.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 B.6(a) or $\{\mathbf{x}_1, \mathbf{x}_2\}$ in Figure B.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* $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$.

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 B.7. Finding coordinates algebraically entails the solution of a system of linear simultaneous equations in which the c_j's are the unknowns:

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

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.[7]
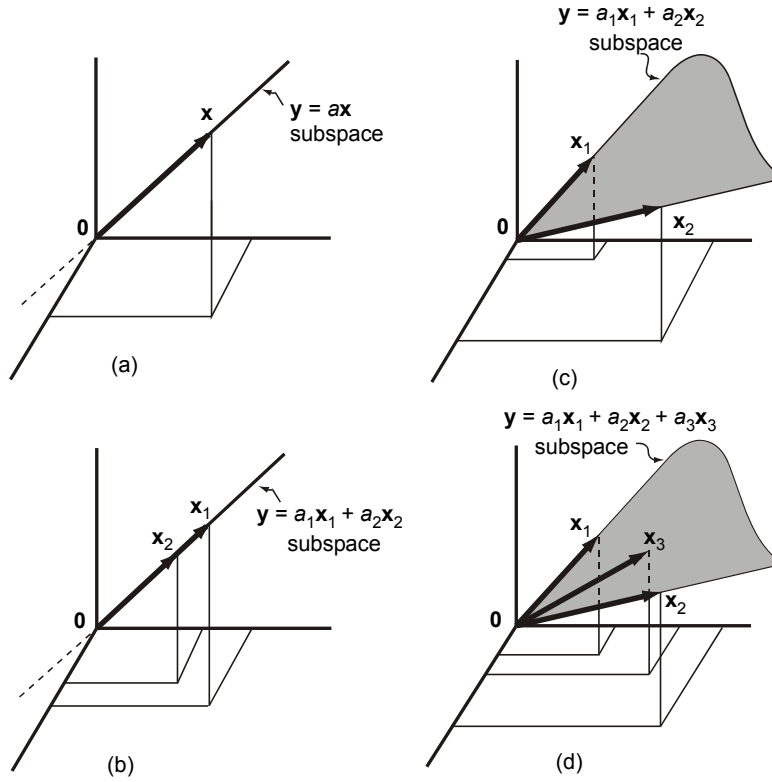
Figure B.6: Subspaces generated by sets of vectors in three-dimensional space. (a) One nonzero vector generates a one-dimensional subspace (a line). (b) Two collinear vectors also generate a one-dimensional subspace. (c) Two linearly independent vectors generate a two-dimensional subspace (a plane). (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.
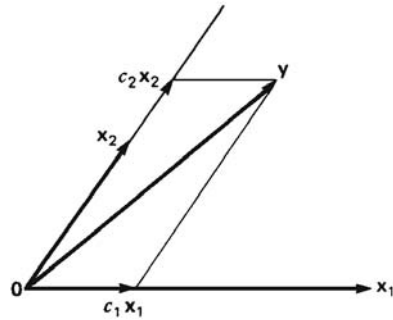
Figure B.7: The coordinates of $\mathbf{y}$ with respect to the basis $\{\mathbf{x}_1, \mathbf{x}_2\}$ of a two-dimensional subspace can be found from the parallelogram rule of vector addition.
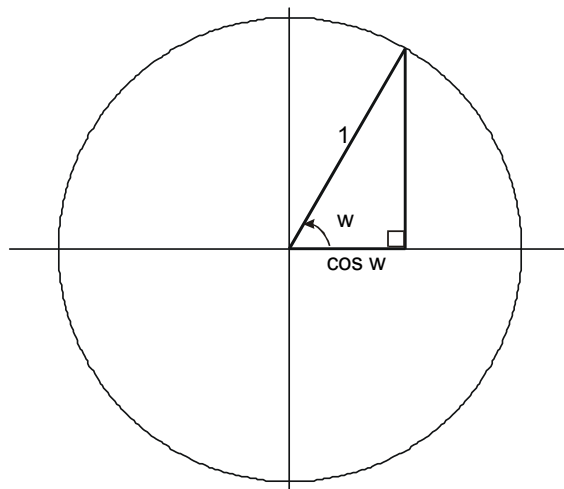
Figure B.8: A unit circle, showing the angle w and its cosine.

B.3.1 Review of Cosines of Angles

Figure B.8 shows a unit circle—that is, a circle of radius 1 centered at the origin. The angle w produces a right triangle inscribed in the circle; notice that the angle is measured in a counter-clockwise direction from the horizontal axis. The cosine of the angle w, denoted $\cos w$, is the signed length of the side of the triangle adjacent to the angle (i.e., "adjacent/hypotenuse," where the hypotenuse is 1 because it is the radius of the unit circle). The cosine function for angles between -360 and 360 degrees is shown in Figure B.9; negative angles represent clockwise rotations. 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.

B.3.2 Orthogonality and Orthogonal Projections

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

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

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 B.10. 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.[8] When $\mathbf{x}$ and $\mathbf{y}$ are orthogonal [as in Figure B.10(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

---

[7] The concept of rank and the solution of systems of linear simultaneous equations are taken up in Section B.4.

[8] I frequently use 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.
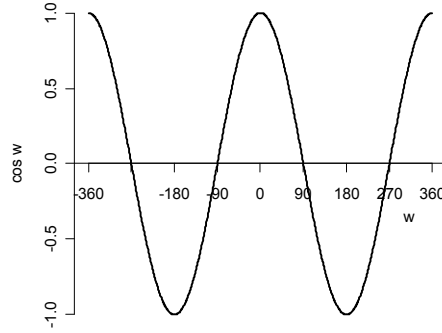
Figure B.9: The cosine function for angles between $w = -360$ and $w = 360$ degrees.

$(\mathbf{x} \cdot \mathbf{x} = \sum x_i^2)$, we have

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

When, in contrast, $\mathbf{x}$ and $\mathbf{y}$ are not orthogonal [as in Figure B.10(b)], then $||\mathbf{x} + \mathbf{y}|| \neq ||\mathbf{x} - \mathbf{y}||$, and $\mathbf{x} \cdot \mathbf{y} \neq 0$.

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.[9] 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 of orthogonal projection is illustrated in Figure B.11. By the Pythagorean theorem (see Figure B.12), $\widehat{\mathbf{y}}$ 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. I will distinguish between two cases:[10] In Figure B.13(a), the angle separating the vectors is between $0°$ and $90°$; in Figure B.13(b), the angle is between $90°$ and $180°$. In the first instance,

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

---

[9]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 $\mathbf{x}_i'\mathbf{x}_i = \mathbf{x}_i \cdot \mathbf{x}_i$.

[10]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 $\cos(360 - w) = \cos w$ (see Figure B.9).
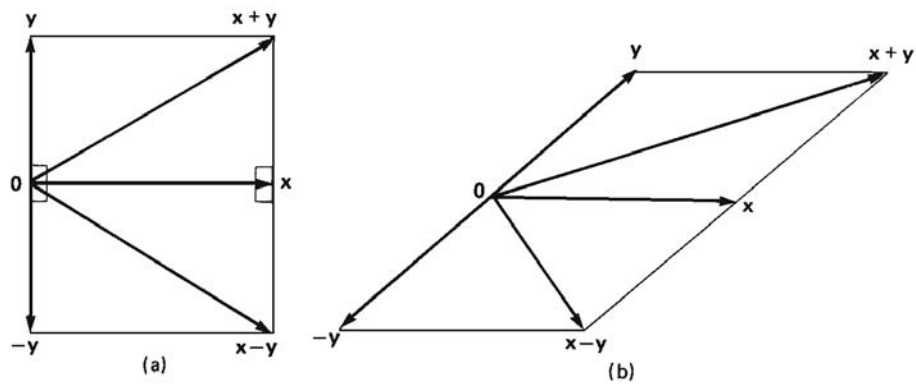
Figure B.10: When two vectors $\mathbf{x}$ and $\mathbf{y}$ are orthogonal, as in (a), their inner product $\mathbf{x} \cdot \mathbf{y}$ is 0. When the vectors are not orthogonal, as in (b), their inner product is nonzero.
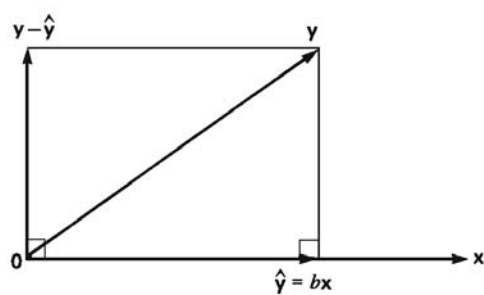

Figure B.11: The orthogonal projection $\widehat{\mathbf{y}} = b\mathbf{x}$ of $\mathbf{y}$ onto $\mathbf{x}$.
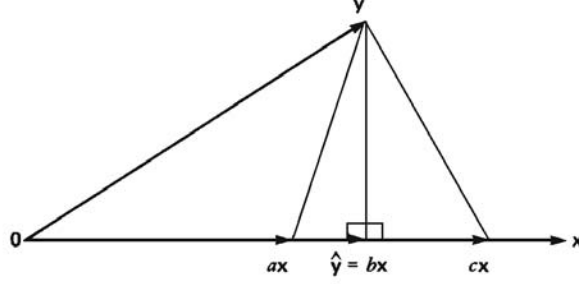
Figure B.12: The orthogonal projection $\widehat{\mathbf{y}} = b\mathbf{x}$ is the point along the line spanned by $\mathbf{x}$ that is closest to $\mathbf{y}$.

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 B.14. 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 $\widehat{\mathbf{y}} = \mathbf{Xb}$. By the definition of an orthogonal projection,

$$\mathbf{x}_j \cdot (\mathbf{y} - \widehat{\mathbf{y}}) = \mathbf{x}_j \cdot (\mathbf{y} - \mathbf{Xb}) = 0 \quad \text{for } j = 1, \ldots, k \tag{B.8}$$

Equivalently, $\mathbf{X}'(\mathbf{y} - \mathbf{Xb}) = \mathbf{0}$, or $\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{Xb}$. 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.

B.4   Matrix Rank and the Solution of Linear Simultaneous Equations

B.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 $\text{rank}(\mathbf{A}) \leq \min(m, n)$.

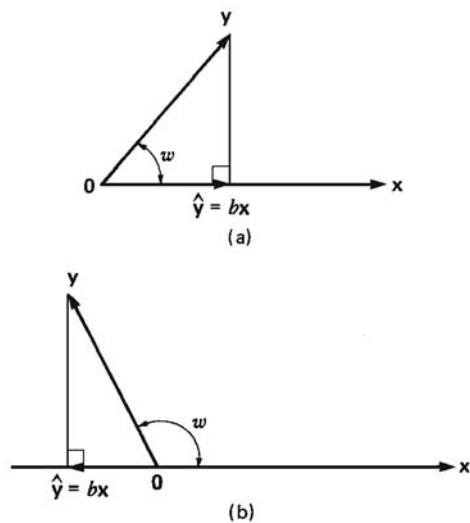Figure B.13: The angle w separating two vectors, $\mathbf{x}$ and $\mathbf{y}$: (a) $0° < w < 90°$; (b) $90° < w < 180°$.
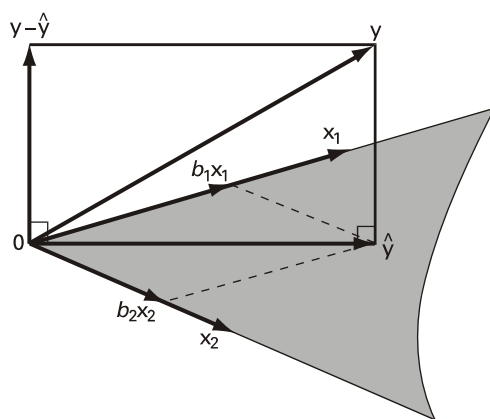

Figure B.14: The orthogonal projection $\widehat{\mathbf{y}}$ of $\mathbf{y}$ onto the subspace (plane) spanned by $\mathbf{x}_1$ and $\mathbf{x}_2$.

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 B.9, where the asterisks represent elements of arbitrary value:

$$
\begin{bmatrix}
0 & \cdots & 0 & 1 & * & \cdots & * & 0 & * & \cdots & * & 0 & * & \cdots & * \\
0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 1 & * & \cdots & * & 0 & * & \cdots & * \\
\vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\
0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 1 & * & \cdots & * \\
0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\
\vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\
0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0
\end{bmatrix}
\begin{array}{l} \text{nonzero} \\ \text{rows} \\ \\ \\ \text{zero} \\ \text{rows} \end{array}
$$

$$(B.9)$$

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 B.1.3. For example, starting with the matrix

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

Divide row 1 by -2,

$$
\begin{bmatrix}
1 & 0 & \frac{1}{2} & -1 \\
4 & 0 & 1 & 0 \\
6 & 0 & 1 & 2
\end{bmatrix}
$$

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

$$
\begin{bmatrix}
1 & 0 & \frac{1}{2} & -1 \\
0 & 0 & -1 & 4 \\
6 & 0 & 1 & 2
\end{bmatrix}
$$

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

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

Multiply row 2 by -1,

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

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

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

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 reduced row-echelon form $\mathbf{A}_R$, 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 $\mathbf{A}$.

B.4.2 Linear Simultaneous Equations

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

$$\underset{(m\times n)}{\mathbf{A}}\ \underset{(n\times 1)}{\mathbf{x}}\ =\ \underset{(m\times 1)}{\mathbf{b}} \tag{B.10}$$

where the elements of the coefficient matrix $\mathbf{A}$ and the right-hand-side vector $\mathbf{b}$ are pre-specified 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 B.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 B.10 produces

$$\mathbf{E}\mathbf{A}\mathbf{x} = \mathbf{E}\mathbf{b} \tag{B.11}$$
$$\mathbf{A}_R\mathbf{x} = \mathbf{b}_R$$

where $\mathbf{b}_R \equiv \mathbf{E}\mathbf{b}$. Equations B.10 and B.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 *over-determined*, 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 *under-determined*.

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 over-determined (if a zero row of $\mathbf{A}_R$ corresponds to a nonzero entry of $\mathbf{b}_R$) or under-determined (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:

Divide row 1 by -2,

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

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

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

Multiply row 2 by -1,

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

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

$$
x_1 + x_4 = \frac{3}{2}
$$
$$
x_3 - 4x_4 = -4
$$
$$
0x_1 + 0x_2 + 0x_3 + 0x_4 = 0
$$

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 the x_1 and x_3 (corresponding to the leading entries) follow:

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

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 are over-determined. If $r < n$, the equations are either over-determined (if inconsistent) or under-determined (if consistent).

To illustrate these results geometrically, consider a system of three linear equations in two unknowns:[11]

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

Each equation describes a line in a two-dimensional coordinate space in which the unknowns define the axes, as illustrated schematically in Figure B.15. If the three lines intersect at a point, as in Figure B.15(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 B.15(b) and (c), then

---

[11]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 B.2.
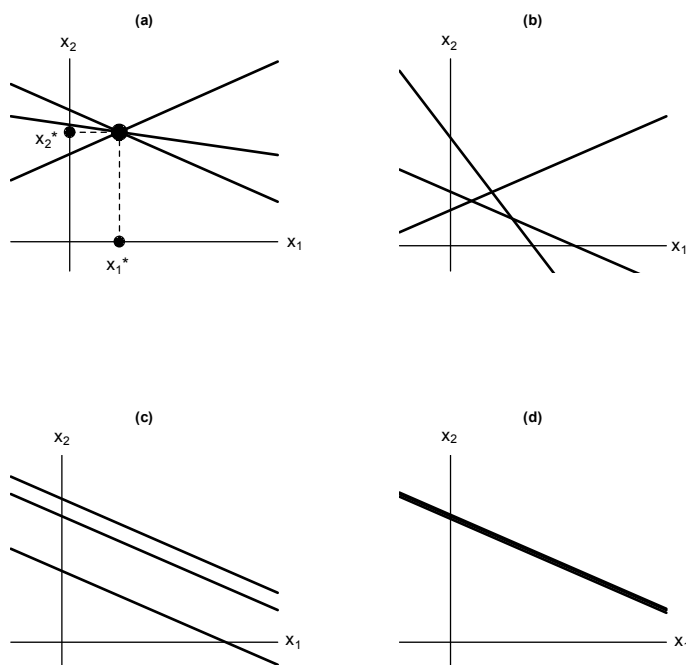
Figure B.15: Three linear equations in two unknowns: (*a*) unique solution; (*b*) and (*c*) over-determined; (*d*) under-determined (three coincident lines).

Table B.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 solution	trivial solution

no pair of values of the unknowns simultaneously satisfies the three equations, which therefore are over-determined. Lastly, if the three lines are coincident, as in Figure B.15(d), then *any* pair of values on the common line satisfies all three equations, and the equations are under-determined.

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

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

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 rank($\mathbf{A}$) $< n$—that is, when the system is under-determined.

The results concerning the solution of linear simultaneous equations developed in this section are summarized in Table B.1.

B.5 Eigenvalues and Eigenvectors

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

$$(\mathbf{A} - l\mathbf{I}_n)\mathbf{x} = \mathbf{0} \tag{B.12}$$

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

$$\det(\mathbf{A} - l\mathbf{I}_n) = 0 \tag{B.13}$$

Equation B.13 is called the *characteristic equation* of the matrix $\mathbf{A}$, and the values of l for which this equation holds are called the *eigenvalues*, *characteristic roots*, or *latent roots* of $\mathbf{A}$. A vector $\mathbf{x}_1$ satisfying Equation B.12 for a particular eigenvalue l_1 is called an *eigenvector*, *characteristic vector*, or *latent vector* of $\mathbf{A}$ associated with l_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

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

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

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

Using the quadratic formula to solve the characteristic equation produces the two roots[12]

$$l_1 = \frac{1}{2}\left[a_{11} + a_{22} + \sqrt{(a_{11} + a_{22})^2 - 4(a_{11}a_{22} - a_{12}a_{21})}\right] \qquad \text{(B.14)}$$

$$l_2 = \frac{1}{2}\left[a_{11} + a_{22} - \sqrt{(a_{11} + a_{22})^2 - 4(a_{11}a_{22} - a_{12}a_{21})}\right]$$

These roots are real if the quantity under the radical is non-negative. Notice, incidentally, that $l_1 + l_2 = a_{11} + a_{22}$ (the sum of the eigenvalues of $\mathbf{A}$ is the trace of $\mathbf{A}$), and that $l_1 l_2 = a_{11}a_{22} - a_{12}a_{21}$ (the product of the eigenvalues is the determinant of $\mathbf{A}$). Furthermore, if $\mathbf{A}$ is singular, then l_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 B.14 becomes

$$l_1 = \frac{1}{2}\left[a_{11} + a_{22} + \sqrt{(a_{11} - a_{22})^2 + 4a_{12}^2}\right] \qquad \text{(B.15)}$$

$$l_2 = \frac{1}{2}\left[a_{11} + a_{22} - \sqrt{(a_{11} - a_{22})^2 + 4a_{12}^2}\right]$$

The eigenvalues of a (2×2) symmetric matrix are necessarily real because the quantity under the radical in Equation B.15 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

$$l_1 = \frac{1}{2}\left[1 + 1 + \sqrt{(1-1)^2 + 4(0.5)^2}\right] = 1.5$$

$$l_2 = \frac{1}{2}\left[1 + 1 - \sqrt{(1-1)^2 + 4(0.5)^2}\right] = 0.5$$

To find the eigenvectors associated with $l_1 = 1.5$, solve the homogeneous system of equations

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

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

---

[12] 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

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

yielding

$$\mathbf{x}_1 = \left[\begin{array}{c} x_{11} \\ x_{21} \end{array} \right] = \left[\begin{array}{c} x_{21}^* \\ x_{21}^* \end{array} \right] = x_{21}^* \left[\begin{array}{c} 1 \\ 1 \end{array} \right]$$

(that is, any vector with two equal entries). Similarly, for $l_2 = 0.5$, solve

$$\left[\begin{array}{cc} 1 - 0.5 & 0.5 \\ 0.5 & 1 - 0.5 \end{array} \right] \left[\begin{array}{c} x_{12} \\ x_{22} \end{array} \right] = \left[\begin{array}{c} 0 \\ 0 \end{array} \right]$$

$$\left[\begin{array}{cc} 0.5 & 0.5 \\ 0.5 & 0.5 \end{array} \right] \left[\begin{array}{c} x_{12} \\ x_{22} \end{array} \right] = \left[\begin{array}{c} 0 \\ 0 \end{array} \right]$$

which produces

$$\mathbf{x}_2 = \left[\begin{array}{c} x_{12} \\ x_{22} \end{array} \right] = \left[\begin{array}{c} -x_{22}^* \\ x_{22}^* \end{array} \right] = x_{22}^* \left[\begin{array}{c} -1 \\ 1 \end{array} \right]$$

(that is, 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, $\det(\mathbf{A} - l\mathbf{I}_n) = 0$, of an $(n \times n)$ matrix is an nth-order polynomial in l; there are, consequently, n eigenvalues, not all necessarily distinct.[13]

- 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.

B.6 Quadratic Forms and Positive-Definite Matrices

The expression

$$\underset{(1 \times n)(n \times n)(n \times 1)}{\mathbf{x}' \quad \mathbf{A} \quad \mathbf{x}} \tag{B.16}$$

---

[13]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.

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 B.16 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.[14]

B.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 (1965), 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.

---

[14]Cf., the geometric discussion following Equation B.8 on page 26.

Appendix C

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—equations of lines and planes, polynomial functions, logarithms, and exponentials—I will briefly take up the following seminal topics in calculus, emphasizing basic concepts: Section C.2, limits of functions; Section C.3, the derivative of a function; Section D.4, the application of derivatives to optimization problems; Section D.5, partial derivatives of functions of several variables, constrained optimization, and differential calculus in matrix form; Section D.6, Taylor series; and Section D.7, the essential ideas of integral calculus.

Although a thorough and rigorous treatment is well beyond the scope of this brief appendix, it is remarkable how far one can get in statistics with a intuitive grounding in the basic ideas of calculus.

C.1 Review

C.1.1 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 C.1, which shows straight lines in the two-dimensional coordinate space with axes x and y; in case case, the line extends infinitely to the left and right beyond the line-segment shown in the graph. When the slope is

---

[1] Newton's claim that Leibniz had appropriated his work touched off one of the most famous priority disputes in the history of science.
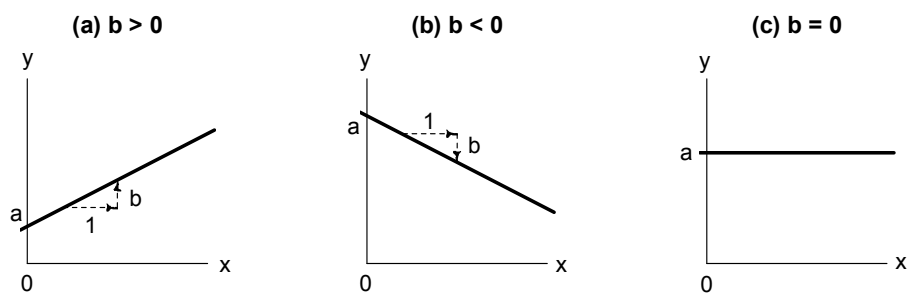
Figure C.1: The graph of a straight line, $y = a + bx$, for (a) $b > 0$, (b) $b < 0$, and (c) $b = 0$.
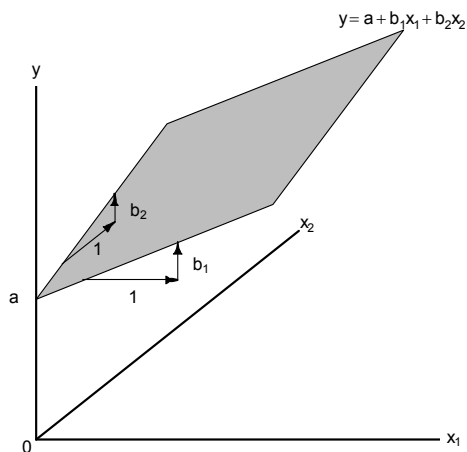
Figure C.2: The equation of a plane, $y = a + b_1 x_1 + b_2 x_2$. Here, both slopes, b_1 and b_2, are positive.

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 3D graph in Figure C.2; the 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.
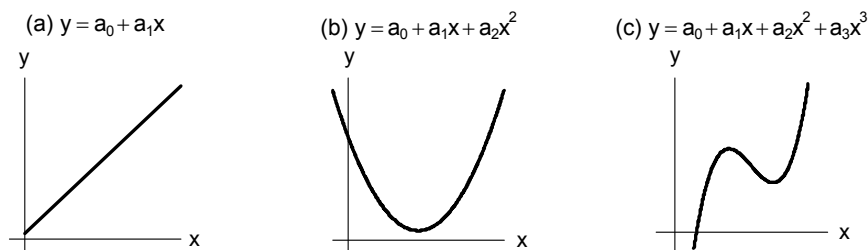
Figure C.3: "Typical" first-order (linear), second-order (quadratic), and third-order (cubic) polynomials.

C.1.2 Polynomials

Polynomials are functions of the form

$$y = a_0 + a_1 x + a_2 x^2 + \cdots + a_p x^p$$

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 C.3, a first-order polynomial is a straight line,

$$y = a_0 + a_1 x$$

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.

C.1.3 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,

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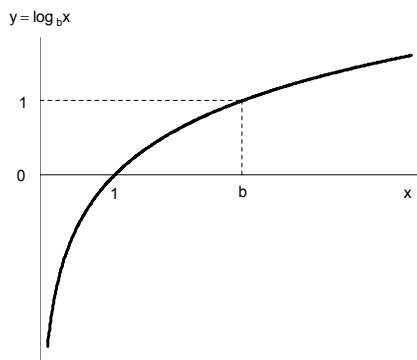
and, similarly,

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

Indeed, the log of 1 to any base is 0, because $b^0 = 1$ for 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 base e are called *natural logs*.[2]

A "typical" log function is graphed in Figure C.1.3.

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

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

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, $y = e^x$, is graphed in Figure C.4.

---

[2] For a justification of this terminology, see Section C.3.4.
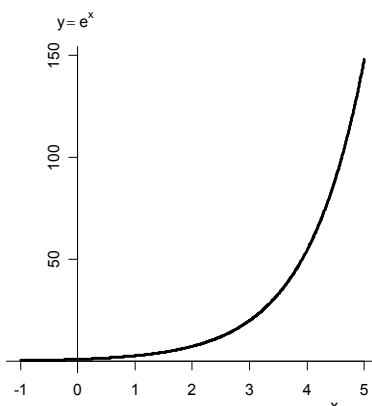
Figure C.4: Graph of the exponential function $y = e^x$.

C.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 *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.

C.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:

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

for all

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

This possibly cryptic definition is clarified by Figure C.5. Note that $f(x_0)$ need not equal L, and need not exist at all. Indeed, limits are often most useful when $f(x)$ does not exist 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."

C.2.2 Finding a Limit: An Example

Find the limit of

$$y = f(x) = \frac{x^2 - 1}{x - 1}$$
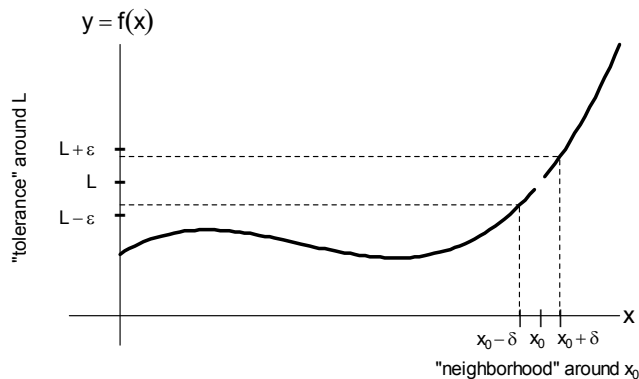
Figure C.5: $\lim_{x \to x_0} f(x) = L$: The limit of the function $f(x)$ as x approaches x_0 is L. The gap in the curve above x_0 is meant to suggest that the function is undefined at $x = x_0$.

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 $x - 1$:

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

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 C.6.

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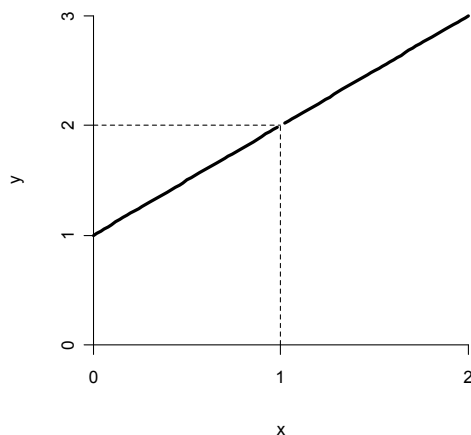
$$\lim_{x \to x_0} f(x) = a$$
$$\lim_{x \to x_0} g(x) = b$$

Figure C.6: $\lim_{x \to 1} \frac{x^2 - 1}{x - 1} = 2$, even though the function is undefined at $x = 1$.

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

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

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

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

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

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

C.3   The Derivative of a Function

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

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

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 C.7, the difference quotient is the slope of the 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
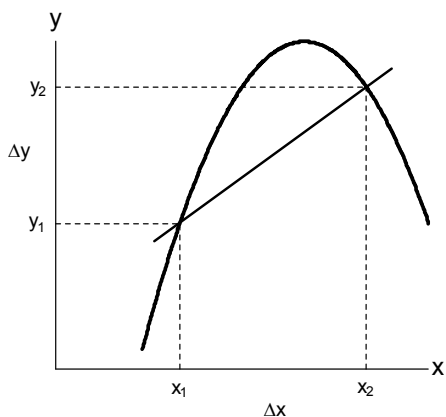
Figure C.7: The difference quotient $\Delta y/\Delta x$ is the slope of the line connecting (x_1, y_1) and (x_2, y_2).

(i.e., as $\Delta x \to 0$):

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

The derivative is therefore the slope of the tangent line to the curve $f(x)$ at $x = x_1$, as shown in Figure C.8.

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.[3] The operation of finding the derivative of a function is called *differentiation*.

C.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:

---

[3] See, e.g., the "chain rule" for differentiation, introduced in Section C.3.3.
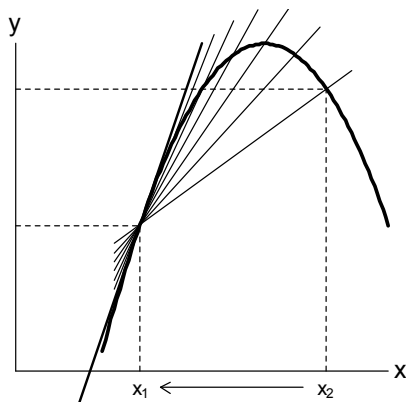
Figure C.8: The derivative is the slope of the tangent line at $f(x_1)$.

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 $f'(x) = 2x = 2 \times 3 = 6$.

C.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

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

C.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 horizontally line in the $\{x, y\}$ plane, and a horizontal line as a slope of 0.

The addition rule, therefore, along with the result that $f'(ax^n) = nax^{n-1}$, serves to differentiate any *polynomial function* (i.e., any weighted sum of powers of x).

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:

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

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

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.

C.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

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

C.3.5   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

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

Notice the alternative notation.

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

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

and so on for *higher-order derivatives*.

For example, the function

$$y = f(x) = 5x^4 + 3x^2 + 6$$
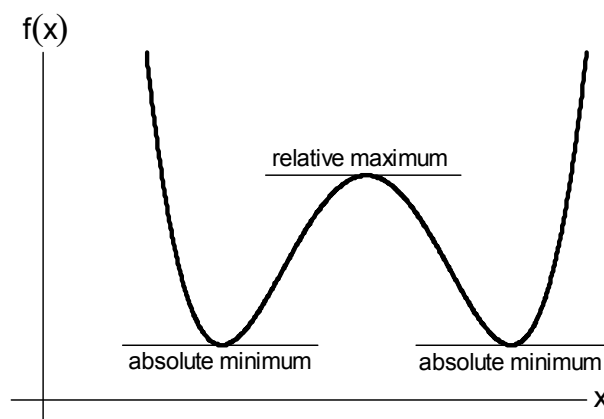
Figure C.9: The derivative (i.e., the slope of the function) is 0 at a minimum or maximum.

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.

C.4 Optimization

An important application of derivatives, both in statistics and more generally, is to *maximization* and *minimization* problems: that is, finding maximum and minimum values of functions (e.g., *maximum* likelihood estimation; *least* squares estimation). Such problems are collectively called *optimization*.

As illustrated in Figure C.9, when a function is at a *relative 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 C.10.

To distinguish among the three cases—minimum, maximum, or neither—we can appeal to the value of the second derivative (see Figure C.11).

- 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.
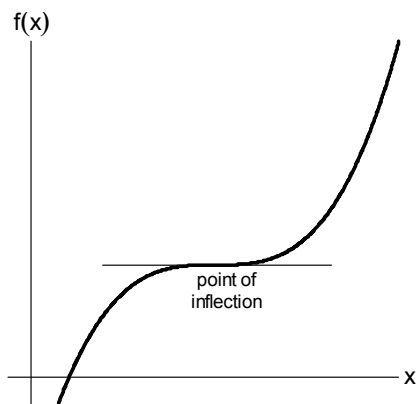
Figure C.10: The derivative is also 0 at a point of inflection.
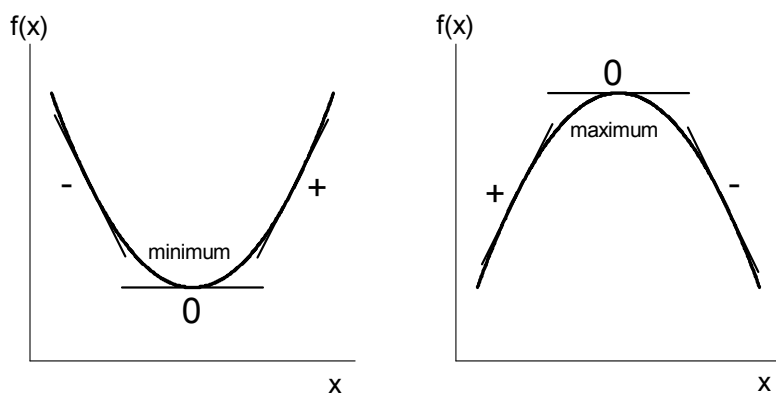
Figure C.11: The first derivative (the slope of the function) is increasing where the function $f(x)$ is at a minimum and decreasing at a maximum.

$y = f(x)$

maximum

minimum   minimum   x

dy/dx

x

d^2y/dx^2

x
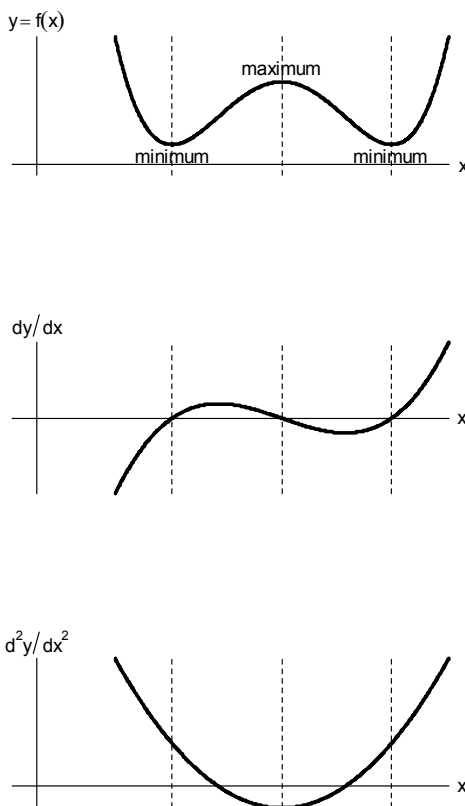
Figure C.12: An example of a function and its first and second derivatives.

- 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*.

The relationships among the original function, the first derivative, and the second derivative are illustrated in Figure C.12: The first derivative is 0 at the two minima and at the (relative) maximum; the second derivative is positive at the two minima, and negative at the maximum.

C.4.1 Optimization: An Example

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

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

The function is shown in Figure C.13. By the way, locating *stationary points*—points at which the first derivative is 0—and determining whether they are minima or maxima (or neither), is helpful in graphing functions.
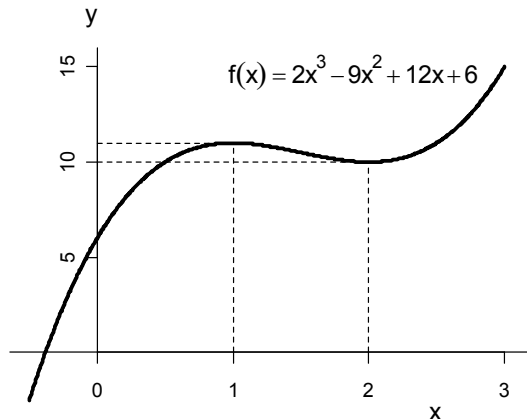
Figure C.13: Finding the extrema of the function $y = f(x) = 2x^3 - 9x^2 + 12x + 6$.

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*.

C.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.

C.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 x's constant. To distinguish it from the ordinary derivative dy/dx, the standard notation for the partial derivative uses Greek deltas in place of d's: $\partial y/\partial x_i$.

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 treat all of the other x's as constants (i.e., literally to hold other x's 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. 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 C.14.

At a local minimum or maximum, the slope of the tangent hyperplane is 0 in all directions. Consequently, to minimize or maximize 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.[5]

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

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

---

[5]I will explain in Section C.5.3 how to distinguish maxima from minima.
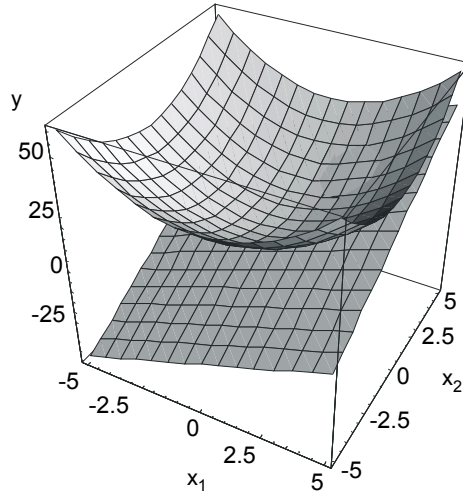
Figure C.14: The function $y = f(x_1, x_2) = x_1^2 + x_1 x_2 + x_2^2 + 10$, showing the tangent plane at $x_1 = 1$, $x_2 = 2$.

The slopes of the tangent plane above the pair of values $x_1 = 1$, $x_2 = 2$, illustrated in Figure C.14, are

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

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

C.5.2   Lagrange Multipliers for Constrained Optimization

The method of Lagrange multipliers[6] 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

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

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

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

---

[6] The method is named after the 18th-century French mathematician J. L. Lagrange.

3. Find the values of $x_1, x_2, \ldots, x_n$ that (along with L) optimize the function $h(x_1,$ $x_2, \ldots, x_n, L)$. That is, differentiate $h(x_1, x_2, \ldots, x_n, L)$ with respect to each of $x_1, x_2, \ldots, x_n$ and L; set the $n+1$ partial derivatives to 0; and solve the resulting system of simultaneous equations for $x_1, x_2, \ldots, x_n$ and L. For the example,

$$\frac{\partial h(x_1, x_2, L)}{\partial x_1} = 2x_1 - L$$

$$\frac{\partial h(x_1, x_2, L)}{\partial x_2} = 2x_2 - L$$

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

Notice that the partial derivative with respect to L, 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 $L = 1$).

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

C.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* 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 $\mathbf{x}$:

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

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

$$y = \underset{(1\times n)(n\times 1)}{\mathbf{a}' \quad \mathbf{x}} = a_1 x_1 + a_2 x_2 + \cdots + a_n x_n$$

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 $\mathbf{x}$,

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

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

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

The partial derivatives are

$$\frac{\partial y}{\partial x_1} = 4x_1 + 6x_2$$
$$\frac{\partial y}{\partial x_2} = 6x_1 + 2x_2$$

And the vector partial derivative is

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

The so-called *Hessian matrix*[7] of second-order partial derivatives of the function $y = f(\mathbf{x})$ is defined in the following manner:

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

For instance, $\partial^2(\mathbf{x}'\mathbf{A}\mathbf{x})/\partial \mathbf{x}\,\partial \mathbf{x}' = 2\mathbf{A}$, for a symmetric matrix $\mathbf{A}$.

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

---

[7] The Hessian is named after the 19th Century German mathematician Ludwig Otto Hesse.

represents a maximum if the Hessian is negative definite.[8]   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,

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

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

C.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*[9]

$$f(x) = f(x_0) + \frac{f'(x_0)}{1!}(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \frac{f'''(x_0)}{3!}(x - x_0)^3 + \cdots \quad \text{(C.1)}$$

$$= \sum_{n=1}^{\infty} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n$$

where $f^{(n)}$ represents the nth-order derivative of f, and $n!$ is the *factorial* of n.[10]

As long as x is sufficiently close to x_0, and as long as the function f 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

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

---

[8] The square matrix $\mathbf{H}$ is *positive definite* if $\mathbf{x}'\mathbf{H}\mathbf{x} > 0$ for any nonzero vector $\mathbf{x}$. (See Section B.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.

[9] Named after the 18th Century British mathematician Brook Taylor.

[10] 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.

Then

$$f'(x) = 2x + 3x^2$$
$$f''(x) = 2 + 6x$$
$$f'''(x) = 6$$
$$f^{(n)}(x) = 0 \text{ for } n > 3$$

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 $x_0 = 2$:

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

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

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

where $\mathbf{g}(\mathbf{x}) \equiv \partial y / \partial \mathbf{x}$ and $\mathbf{H}(\mathbf{x}) = \partial^2 y / \partial \mathbf{x} \, \partial \mathbf{x}'$ 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 C.1.

C.7   Essential Ideas of Integral Calculus

C.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 C.15. This area can be approximated by dividing the line 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 C.16. The x-values defining the rectangles are

$$x_0, x_0 + \Delta x, x_0 + 2\Delta x, \ldots, x_0 + n\Delta x$$
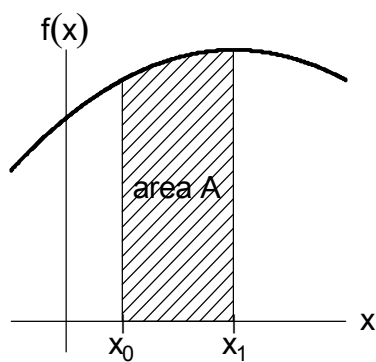
Figure C.15: The area A under a function $f(x)$ between x_0 and x_1.
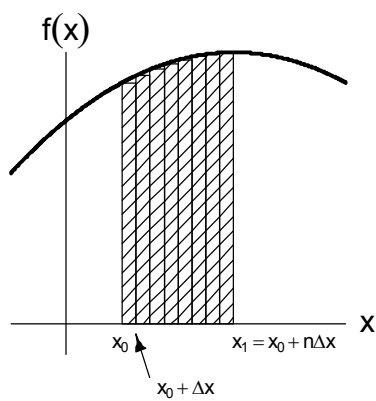
Figure C.16: Approximating the area under a curve by summing the areas of rectangles.
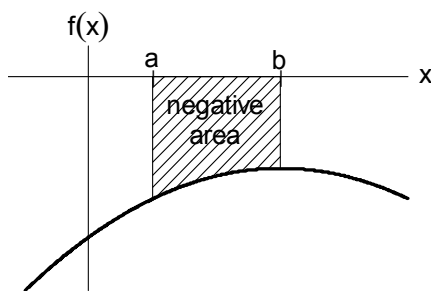
Figure C.17: The integral $\int_a^b f(x)dx$ is negative because the y values are negative between the limits of integration a and b.

Consequently the combined area of the rectangles is

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

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

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

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 C.17.

C.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 $f(x)$.
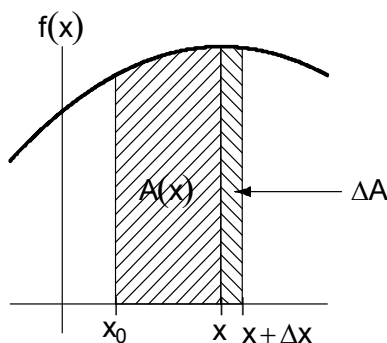
Figure C.18: The area $A(x)$ under the curve between the fixed value x_0 and another value x.

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 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.

C.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 non-rigorous 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 C.18. 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 C.18, $x + \Delta x$ 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

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

The area ΔA is also

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

Taking the derivative of A,

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

Consequently,

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

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,

$$A(x_0) = F(x_0) + c = 0$$
$$c = -F(x_0)$$

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)

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

It is convenient to use[11]

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

Then

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

---

[11]Reader: Verify that $F(x)$ is an antiderivative of $f(x) = x^2 + 3$.

C.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, see Binmore (1983).

Appendix D

Probability and Estimation

The purpose of this appendix is to outline basic results in probability and statistical inference that are employed principally in the starred parts of the text. Material in the un-starred portions of this appendix is, however, used occasionally in the un-starred parts of the text. For good reason, elementary statistics courses—particularly in the social sciences—often provide only the barest introduction to probability and the theory of estimation. After a certain point, however, some background in these topics is necessary.

In Section D.1, I review concepts in elementary probability theory. Sections D.2 and D.3 briefly describe a number of probability distributions that are of special importance in the study of linear and related models. Section D.4 outlines asymptotic distribution theory, which we occasionally require to determine properties of statistical estimators, a subject that is taken up in Section D.5. Section D.6, develops the broadly applicable and centrally important method of maximum-likelihood estimation. The concluding section of the appendix, Section D.7, introduces Bayesian estimation. Taken together, the sections of this appendix provide a "crash course" in some of the basics of mathematical statistics.

D.1 Elementary Probability Theory

D.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 (i.e., countably infinite[1]), 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, $S = \{1, 2, 3, \ldots\}$. If we burn a light bulb until it fails, recording the burning time in

---

[1]To say that a set is countably infinite means that a one-to-one relationship can be established between the elements of the set and the natural numbers 0, 1, 2,....

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.

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.[2] *Probabilities* are numbers assigned to events in a manner consistent with the following axioms (rules):

P1: $\Pr(E) \geq 0$: The probability of an event is nonnegative.

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 $\Pr(\emptyset) = 0$, where $\emptyset$ is the *empty event*, which contains no outcomes: 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,[3] so that

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

Then, for $E \equiv \{HH, HT\}$, $\Pr(E) = .25 + .25 = .5$.

In *classical statistics*, the perspective adopted in most applications of statistics (and, with few exceptions, in the body of the text), 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—that is, probabilities are interpreted as long-run relative frequencies (proportions).[4]

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 $E_2 \cap E_3 = \{TH\}$.

---

[2] The subscripts $a, b, \ldots, m$ are each (different) numbers between 1 and n.

[3] 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.

[4] Cf., Section D.8 on Bayesian statistical inference.

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

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

The conditional probability is interpreted as the probability that E_2 will occur if E_1 is known to have occurred. Two events are *independent* if $\Pr(E_1 \cap E_2) = \Pr(E_1)\Pr(E_2)$.[5] 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

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

---

[5]Independence is different from disjointness: If two events are disjoint, then they cannot occur together, and they are, therefore, *dependent*.

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.

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,

$$\Pr(E_1 \cap E_2 \cap \cdots \cap E_k) = \Pr(E) = 1 - \Pr(\overline{E}) \tag{D.1}$$

$$\geq 1 - \sum_{i=1}^{k} \Pr(\overline{E}_i)$$

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

$$\Pr(E_1 \cap E_2 \cap \cdots \cap E_k) \equiv 1 - a \tag{D.2}$$

$$\geq 1 - kb$$

Equation D.2 and the more general Equation D.1 are called *Bonferroni inequalities.*[6]

Equation D.2 has the following application to simultaneous statistical inference: Suppose that b is the Type I error rate for each of k non-independent 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 "data dredging" can prove seriously misleading.

D.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. The probabilities $p(x)$ for all values of X comprise the *probability distribution* of the random variable. If, for example, each of the four outcomes of the coin-flipping experiment has probability .25,

---

[6] The Bonferroni inequalities are named after Carlo Emilio Bonferroni, a 20th Century Italian mathematician.

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 outcomes 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:

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

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 $\Pr(X \le x)$, but it generally becomes meaningless to refer to the probability of observing individual values of X. The *probability density function* $p(x)$ is, nevertheless, the continuous analog of the discrete probability distribution, defining $p(x) \equiv dP(x)/dx$. Reversing this relation,[7] $P(x) = \int_{-\infty}^{x} p(x)\, dx$; and

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

Thus, as illustrated in Figure D.1, areas under the density function are interpreted as probabilities.

A particularly simple continuous probability distribution is the *rectangular distribution*:

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

This density function is pictured in Figure D.2(a), and the corresponding cumulative distribution function is shown in Figure D.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 $a \le X \le b$.

---

[7]If you are unfamiliar with integral calculus (which is described in Section C.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.
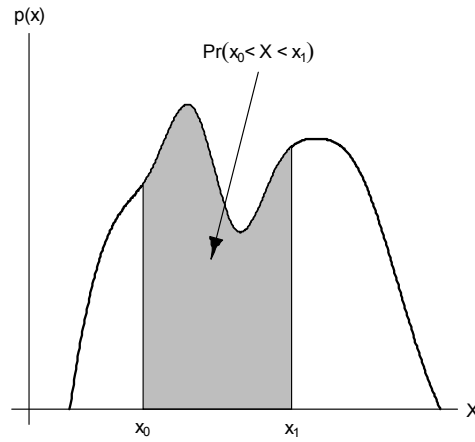
Figure D.1: Areas under the probability density function $p(x)$ are interpreted as probabilities.
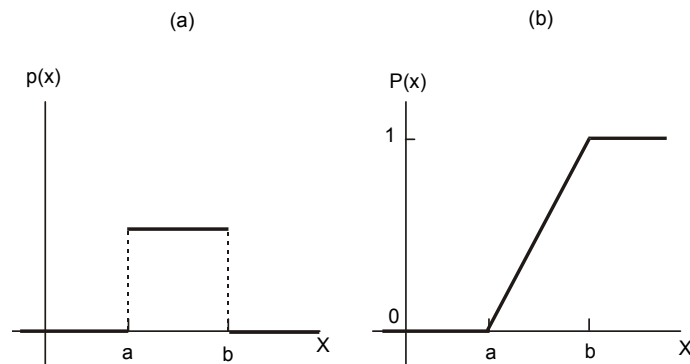

Figure D.2: (a) The probability density function $p(x)$, and (b) the cumulative distribution function $P(x)$ for the rectangular distribution.

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.

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

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

The analogous formula for the continuous case is

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

The variance of a random variable X, written $V(X)$ or σ_X^2, is defined as $E[(X - \mu_X)^2]$. Thus, in the discrete case,

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

while, in the continuous case,

$$V(X) \equiv \int_{-\infty}^{\infty} (x - \mu_X)^2 p(x) \, dx$$

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 $\sigma = \sqrt{0.5} \approx 0.707$.

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 $p(x_1)$.

---

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

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, $p(x_1|x_2)$ is

$p(x_1|x_2)$		
	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 $V(X_1|x_2)$. Using the illustrative conditional distributions $p(x_1|x_2)$,

$$E(X_1|0) = 0(0) + 1(1) + 0(2) = 1$$
$$V(X_1|0) = 0(0 - 1)^2 + 1(1 - 1)^2 + 0(2 - 1)^2 = 0$$
$$E(X_1|1) = .5(0) + 0(1) + .5(2) = 1$$
$$V(X_1|1) = .5(0 - 1)^2 + 0(1 - 1)^2 + .5(2 - 1)^2 = 1$$

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:

$$C(X_1, X_2) = \sigma_{12} \equiv E[(X_1 - \mu_1)(X_2 - \mu_2)]$$

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. 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:

$$E(\mathbf{x}) = \boldsymbol{\mu}_x \equiv [E(X_1), E(X_2), \ldots, E(X_n)]'$$

The *variance-covariance matrix* of a vector random variable $\mathbf{x}$ is defined in analogy to the scalar variance as

$$V(\mathbf{x}) = \underset{(n \times n)}{\boldsymbol{\Sigma}_{xx}} \equiv E[(\mathbf{x} - \boldsymbol{\mu}_x)(\mathbf{x} - \boldsymbol{\mu}_x)'] = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{bmatrix}$$

The diagonal entries of $V(\mathbf{x})$ are the variances of the X's, and the off-diagonal entries are their covariances. The variance-covariance matrix $V(\mathbf{x})$ is symmetric and positive semi-definite.[9] The *covariance matrix* of two vector random variables $\underset{(n \times 1)}{\mathbf{x}}$ and $\underset{(m \times 1)}{\mathbf{y}}$ is

$$C(\mathbf{x}, \mathbf{y}) = \underset{(n \times m)}{\boldsymbol{\Sigma}_{xy}} \equiv E[(\mathbf{x} - \boldsymbol{\mu}_x)(\mathbf{y} - \boldsymbol{\mu}_y)']$$

and consists of the covariances between pairs of X's and Y's.

D.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

$$E(Y) = \mu_Y = \sum_x (a + bx)p(x)$$

$$= a \sum p(x) + b \sum xp(x)$$

$$= a + b\mu_X$$

---

[9] See Section B.6.

and (employing this property of the expectation operator)

$$V(Y) = E[(Y - \mu_Y)^2] = E\{[(a + bX) - (a + b\mu_X)]^2\}$$
$$= b^2 E[(X - \mu_X)^2] = b^2 \sigma_X^2$$

Now, let Y be a linear function $a_1 X_1 + a_2 X_2$ 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

$$E(Y) = \mu_Y = \sum_{x_1} \sum_{x_2} (a_1 x_1 + a_2 x_2) p(x_1, x_2)$$
$$= \sum_{x_1} \sum_{x_2} a_1 x_1 p(x_1, x_2) + \sum_{x_1} \sum_{x_2} a_2 x_2 p(x_1, x_2)$$
$$= a_1 \sum_{x_1} x_1 p(x_1) + a_2 \sum_{x_2} x_2 p(x_2)$$
$$= a_1 \mu_1 + a_2 \mu_2$$

and

$$V(Y) = E[(Y - \mu_Y)^2]$$
$$= E\{[(a_1 X_1 + a_2 X_2) - (a_1 \mu_1 + a_2 \mu_2)]^2\}$$
$$= a_1^2 E[(X_1 - \mu_1)^2] + a_2^2 E[(X_2 - \mu_2)^2]$$
$$\quad + 2a_1 a_2 E[(X_1 - \mu_1)(X_2 - \mu_2)]$$
$$= a_1^2 \sigma_1^2 + a_2^2 \sigma_2^2 + 2a_1 a_2 \sigma_{12}$$

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]

$$E(Y) = \int_{-\infty}^{\infty} (a + bx) p(x)\, dx$$
$$= a \int_{-\infty}^{\infty} p(x)\, dx + b \int_{-\infty}^{\infty} x p(x)\, dx$$
$$= a + bE(X)$$

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 $\underset{(m \times n)}{\mathbf{A}} \underset{(n \times 1)}{\mathbf{x}}$ of the vector random variable $\mathbf{x}$, which has expectation $E(\mathbf{x}) = \boldsymbol{\mu}_x$ and variance-covariance matrix $V(\mathbf{x}) = \boldsymbol{\Sigma}_{xx}$. Then it can be shown (in a manner analogous to the scalar proofs given previously) that

$$E(\mathbf{y}) = \boldsymbol{\mu}_y = \mathbf{A} \boldsymbol{\mu}_x$$
$$V(\mathbf{y}) = \boldsymbol{\Sigma}_{yy} = \mathbf{A} \boldsymbol{\Sigma}_{xx} \mathbf{A}'$$

---

[10]If you are unfamiliar with calculus, then simply think of the integral $\int$ as the continuous analog of the sum $\sum$.

If the entries of $\mathbf{x}$ are pair-wise independent, then all of the off-diagonal entries of $\boldsymbol{\Sigma}_{xx}$ are 0, and the variance of each entry of $\mathbf{y}$ takes an especially simple form:

$$\sigma_{Y_i}^2 = \sum_{j=1}^{n} a_{ij}^2 \sigma_{X_j}^2$$

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,[11] is the absolute value of the $(n \times n)$ determinant

$$\det \begin{bmatrix} \dfrac{\partial X_1}{\partial Y_1} & \cdots & \dfrac{\partial X_n}{\partial Y_1} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial X_1}{\partial Y_n} & \cdots & \dfrac{\partial X_n}{\partial Y_n} \end{bmatrix}$$

and $|\det(\partial \mathbf{y}/\partial \mathbf{x})|$ is similarly defined.

D.2   Some Discrete Probability Distributions

In this section, I define four important families of discrete probability distributions: the binomial distributions; their generalization, 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*."

D.2.1   The Binomial Distributions

The coin-flipping experiment described at the beginning of Section D.1.2 gives rise to a binomial random variable that counts the number of heads in two 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. The probability of observing exactly x heads and $n - x$ tails [i.e., $\Pr(X = x)$] is given by the *binomial distribution*:

$$p(x) = \binom{n}{x} \pi^x (1 - \pi)^{n-x} \tag{D.3}$$

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 $\binom{n}{x} \equiv$

---

[11]The Jacobian is named after the 19th Century German mathematician Carl Gustav Jacob Jacobi.
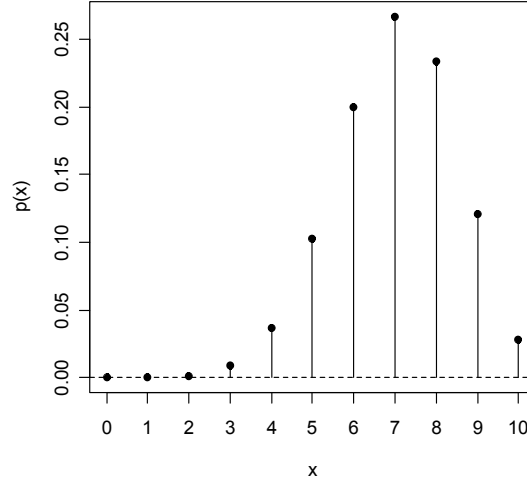
Figure D.3: The binomial distribution for $n = 10$ and $\pi = .7$.

$n!/[x!(n-x)!]$, 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 D.3 shows the binomial distribution for $n = 10$ and $\pi = .7$.

D.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 1's, X_2 the number of 2's, ..., X_6 the number of 6's. 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

---

[12]The exclamation point is the *factorial* operator:

$$n! \equiv n \times (n-1) \times \cdots \times 2 \times 1 \text{ for integer } n > 1$$
$$\equiv 1 \text{ for } n = 0 \text{ or } 1$$

particular arrangement; and $n!/(x_1!x_2!\cdots x_k!)$ counts the number of *different* arrangements. Finally, if $k = 2$, then $x_2 = n - x_1$, and the multinomial distribution reduces to the binomial distribution of Equation D.3.

D.2.3 The Poisson Distributions

The 19th century French mathematician S. 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

- 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 non-overlapping subintervals is independent.

The expectation of a Poisson random variable is $E(X) = \lambda$, and its variance is also $V(X) = \lambda$. Figure D.4 illustrates the Poisson distribution with rate parameter $\lambda = 5$ (implying that, on average, five events occur during the fixed period of observation).

D.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 counts the number of *tails* that are observed before the target is reached. Then X follows a *negative binomial distribution*, with probability function

$$p(x) = \binom{s + x - 1}{x} \pi^s (1 - \pi)^x \text{ for } x = 0, 1, 2, \ldots$$

where π is the probability of a head on an individual flip of the coin. The expectation of the negative binomial random variable is $E(X) = s(1 - \pi)/\pi$, and its variance is $V(X) = s(1 - \pi)/\pi^2$. Figure D.5 shows the negative binomial distribution for $s = 4$ and $\pi = .5$.

An alternative route to the negative binomial distribution is as a mixture of Poisson random variables whose means follow a gamma distribution with scale parameter $(1 - \pi)/\pi$ and shape parameter s (in which case s need not be an integer).[13]

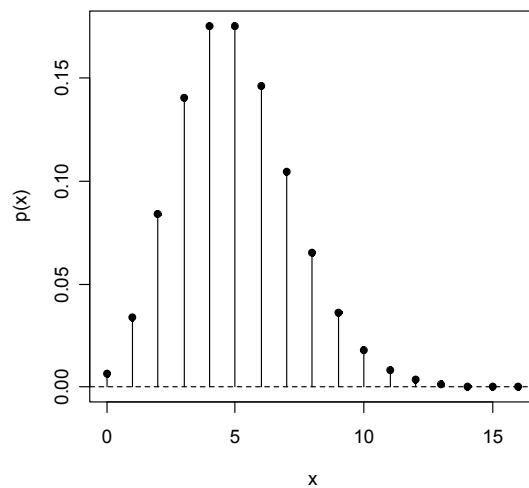[13]The gamma distribution is described in Section D.3.7.

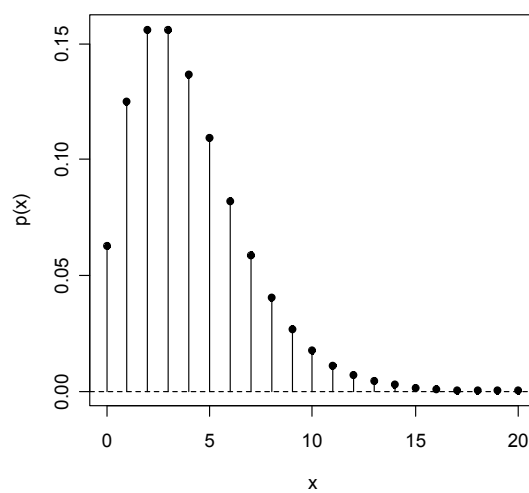Figure D.4: The Poisson distribution with rate parameter $\lambda = 5$.

Figure D.5: Negative binomial distribution for $s = 4$ and $\pi = .5$.
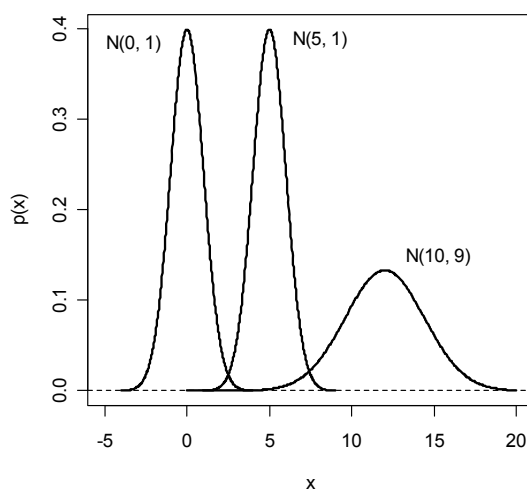
Figure D.6: Normal density functions: $N(0, 1)$, $N(5, 1)$, and $N(10, 9)$.

D.3   Some Continuous Distributions

In this section, I describe five families of continuous random variables that play central roles in the development of linear statistical models: the univariate normal, chi-square, t-, and F-distributions, and the multivariate-normal distribution. I also describe the inverse Gaussian, gamma, and beta distributions—the first two of these because of their role in generalized linear models (the subject of Chapter 15), and the last because of its use in Section D.8 on Bayesian statistical inference. Despite the relatively complex formulas defining the continuous distributions in this section, I have left most of the section un-starred, because some familiarity with the normal, chi-square, t-, and F-distributions is important to understanding statistical inference in linear models.[14]

D.3.1   The Normal Distributions

A *normally distributed* (or *Gaussian*[15]) random variable X has probability density function

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x - \mu)^2}{2\sigma^2}\right]$$

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 D.6. I frequently use the abbreviated notation $X \sim N(\mu, \sigma^2)$, meaning that X is normally distributed with expectation μ and variance σ^2.

---

[14] You may, if you wish, skip the formulas in favor of the graphs and verbal descriptions of the several distributions.

[15] The Gaussian distributions are named after the great German mathematician Carl Friedrich Gauss (1777–1855), although they were first introduced in 1734 by the French mathematician Abraham de Moivre as an approximation to the binomial distribution.
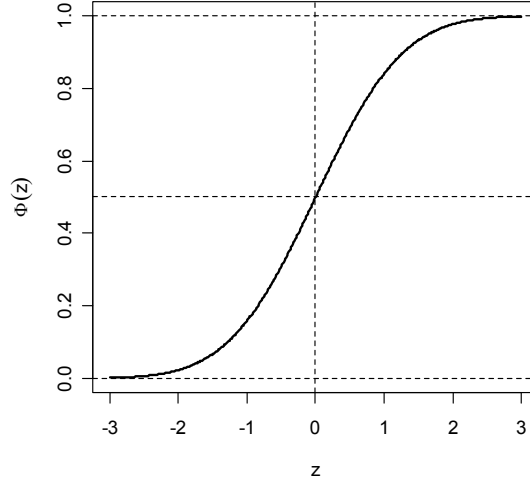
Figure D.7: The CDF of the unit-normal distribution, $\Phi(z)$.

Of particular importance is the *unit-normal* random variable $Z \sim N(0, 1)$, with density function

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp(-z^2/2)$$

The CDF of the unit-normal distribution, $\Phi(z)$, is shown in Figure D.7.

D.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

$$P(x^2) = \frac{1}{2^{n/2}\Gamma(\frac{n}{2})}(x^2)^{(n-2)/2} \exp(-x^2/2)$$

where $\Gamma(\cdot)$ is the *gamma function*[*]

$$\Gamma(x) = \int_0^\infty e^{-z} z^{x-1} dz \tag{D.4}$$

(for the generic argument x), which is a kind of continuous generalization of the factorial function; in particular, when x is a non-negative integer, $x! = \Gamma(x + 1)$. In the current case,

$$\Gamma\left(\frac{n}{2}\right) \equiv \begin{cases} \left(\frac{n}{2} - 1\right)! & \text{for } n \text{ even} \\ \left(\frac{n}{2} - 1\right)\left(\frac{n}{2} - 2\right) \cdots \left(\frac{3}{2}\right)\left(\frac{1}{2}\right)\sqrt{\pi} & \text{for } n \text{ odd} \end{cases}$$

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 D.8.
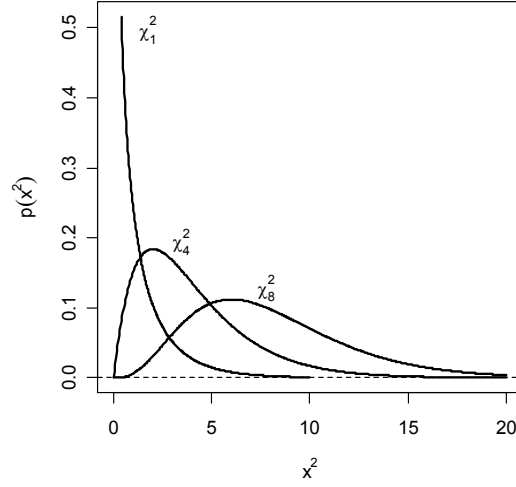
Figure D.8: Chi-square density functions: χ_1^2, χ_4^2, and χ_8^2.

D.3.3 The t-Distributions

If Z follows a unit-normal distribution, and X^2 independently follows a chi-square distribution with n degrees of freedom, then

$$t \equiv \frac{Z}{\sqrt{\dfrac{X^2}{n}}}$$

is a t random variable with n degrees of freedom, abbreviated t_n.[16] The probability density function of t is

$$p(t) = \frac{\Gamma\left(\dfrac{n+1}{2}\right)}{\sqrt{\pi n}\,\Gamma\left(\dfrac{n}{2}\right)} \times \frac{1}{\left(1 + \dfrac{t^2}{n}\right)^{(n+1)/2}} \tag{D.5}$$

From the symmetry of this formula around $t = 0$, it is clear that $E(t) = 0$.[17] 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 D.9. As degrees of freedom grow, the t-distribution more and more closely approximates 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.

---

[16] I write a lowercase t for the random variable in deference to nearly universal usage.

[17] 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.
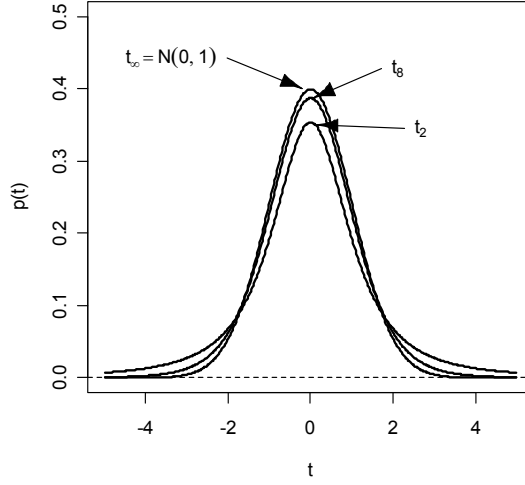
Figure D.9: t density functions: t_2, t_8, and $N(0, 1) = t_\infty$.

D.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

$$F \equiv \frac{X_1^2/n_1}{X_2^2/n_2}$$

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 probability density for F is

$$p(f) = \frac{\Gamma\left(\dfrac{n_1 + n_2}{2}\right)}{\Gamma\left(\dfrac{n_1}{2}\right)\Gamma\left(\dfrac{n_2}{2}\right)} \left(\frac{n_1}{n_2}\right)^{n_1/2} f^{(n_1-2)/2} \left(1 + \frac{n_1}{n_2}f\right)^{-(n_1+n_2)/2} \tag{D.6}$$

Comparing Equations D.5 and D.6, it can be shown that $t_n^2 = F_{1, n}$. As n_2 grows larger, F_{n_1, n_2} approaches $\chi_{n_1}^2/n_1$ and, in the limit, $F_{n, \infty} = \chi_n^2/n$.

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 D.10 shows several F probability density functions.

D.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

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}\sqrt{\det \boldsymbol{\Sigma}}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right]$$
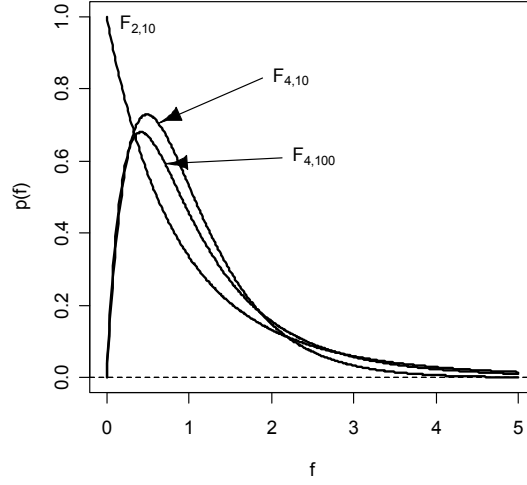
Figure D.10: F density functions: $F_{2,\,10}$, $F_{4,\,10}$, and $F_{4,\,100}$.

which I abbreviate 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, $X_i \sim N(\mu_i, \sigma_i^2)$;[18] and the conditional distributions 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 normal. Furthermore, if $\mathbf{x} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and

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

is a linear transformation of $\mathbf{x}$ with rank$(\mathbf{A}) = m \leq n$, then $\mathbf{y} \sim N_m(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}A')$. 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 D.11.

D.3.6 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 D.12 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.

---

[18]The converse is *not* true: Each X_i can be *univariately* normally distributed without $\mathbf{x}$ being *multivariate* normal.
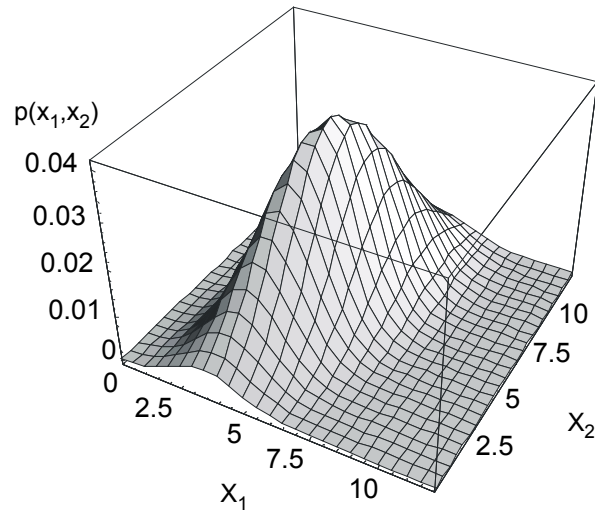
Figure D.11: The bivariate-normal density function for $\mu_1 = 5$, $\mu_2 = 6$, $\sigma_1 = 1.5$, $\sigma_2 = 3$, and $\sigma_{12} = 2.25$. The slices of the density surface (representing the conditional distributions of each variable given values of the other) are normal both in the direction of X_1 and in the direction of X_2.
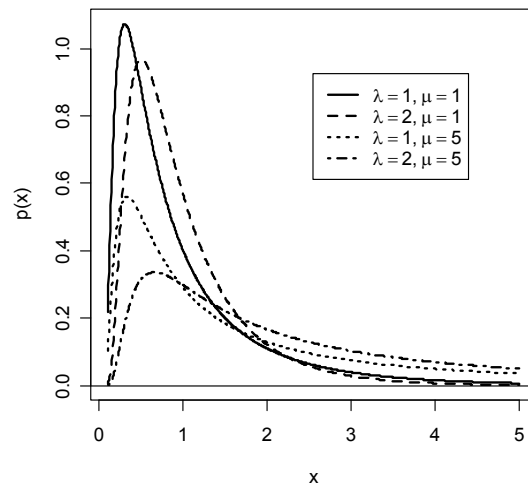
Figure D.12: Inverse-Gaussian distributions for several combinations of values of the parameters μ and λ.
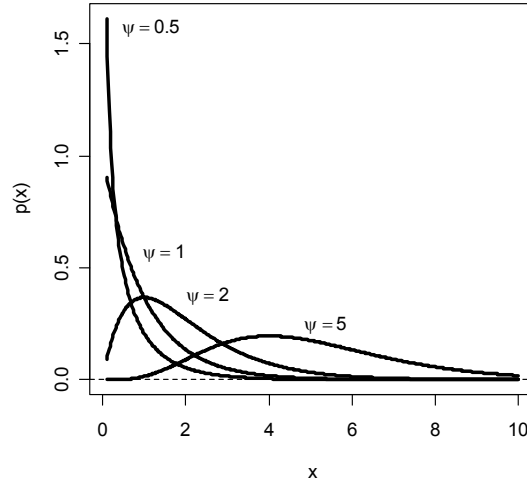
D.3.7   The Gamma Distributions*

The gamma distributions are a family of continuous distributions with probability-density function indexed by the *scale parameter* $\omega > 0$ and *shape parameter* $\psi > 0$:

$$p(x) = \left(\frac{x}{\omega}\right)^{\psi-1} \times \frac{\exp\left(\dfrac{-x}{\omega}\right)}{\omega\Gamma(\psi)} \text{ for } x > 0$$

where $\Gamma(\cdot)$ is the gamma function.[19]   The expectation and variance of the gamma distribution are, respectively, $E(X) = \omega\psi$ and $V(X) = \omega^2\psi$. Figure D.3.7 shows gamma distributions for scale $\omega = 1$ and several values of the shape ψ. (Altering the scale parameter would change only the labelling of the horizontal axis in the graph.) As the shape parameter gets larger, the distribution grows more symmetric.

D.3.8   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

$$V(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

---

[19]See Equation D.4 (on page 80).
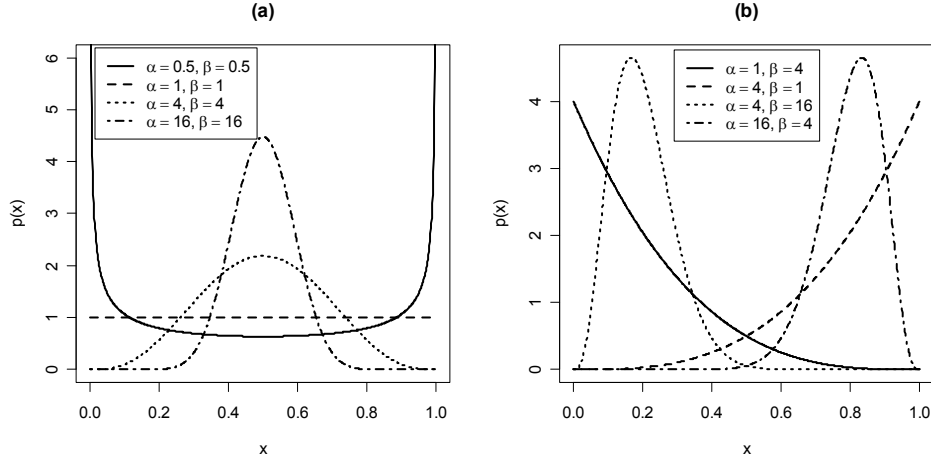
Figure D.13: Beta distributions for several combinations of values of the scale parameters α and β. As is apparent in panel (a), the beta distribution reduces to the rectangular distribution when $\alpha = \beta = 1$. Beta distributions for several combinations of values of the scale parameters α and 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).

The expectation, therefore, depends upon the relative size of the parameters, with $E(X) = 0.5$ when $\alpha = \beta$. The skewness of the beta distribution also depends upon the relative sizes of the parameters, and the distribution is symmetric when $\alpha = \beta$. The variance declines as α and β grow. Figure D.13 shows several beta distributions. As is apparent from these graphs, the shape of the beta distribution is very flexible.

D.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 appendix.

D.4.1 Probability Limits

Although asymptotic distribution theory applies to sequences of random variables, it is necessary first to consider the *non-stochastic infinite sequence* $\{a_1, a_2, \ldots, a_n, \ldots\}$.[20] 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 $n(\epsilon)$ such that $|a_n - a| < \epsilon$ for all $n > n(\epsilon)$. In words: a_n can be made arbitrarily close to a by picking n sufficiently

---

[20]By "non-stochastic" I mean that each a_n is a fixed number rather than a random variable.
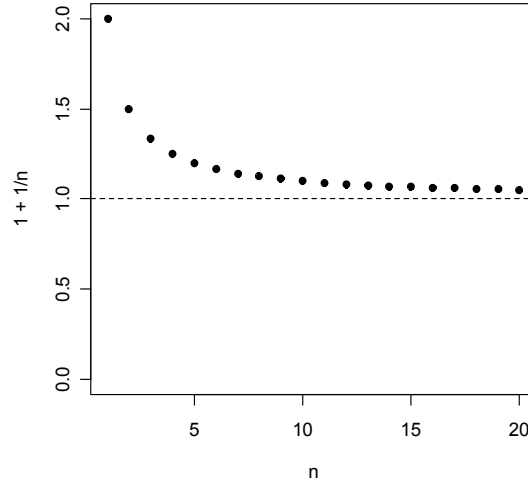
Figure D.14: The first 20 values of the sequence $a_n = 1 + 1/n$, which has the limit $a = 1$.

large.[21]   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 D.14.

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 *non-stochastic* sequence of probabilities $\{p_1, p_2, \ldots, p_n, \ldots\}$ approaches a limit of 1;[22] that is, $\lim_{n \to \infty} \Pr(|X_n - a| < \delta) = 1$. Then, as n grows, the random variable X_n concentrates more and more of its probability in a small region around a, a situation that is illustrated in Figure D.15. If this result holds regardless of how small δ is, then we say that a is the *probability limit* of X_n, denoted $\text{plim} \, X_n = a$. We generally drop the subscript n to write the even more compact expression, $\text{plim} \, X = a$.

Probability limits have the following very useful property: If $\text{plim} \, X = a$, and if $Y = f(X)$ is some continuous function of X, then $\text{plim} \, Y = f(a)$. Likewise, if $\text{plim} \, X = a$, $\text{plim} \, Y = b$, and $Z = f(X, Y)$ is a continuous function of X and Y, then $\text{plim} \, Z = f(a, b)$.

D.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 $\mathcal{E}(X)$.

---

[21]The notation $n(\epsilon)$ 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 C.2.

[22]To say that $\{p_1, p_2, \ldots, p_n, \ldots\}$ is a *non-stochastic* 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.
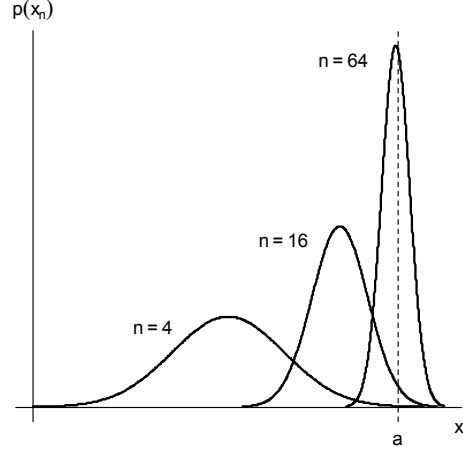
Figure D.15: $\operatorname{plim} X_n = a$: As n grows, the distribution of X_n concentrates more and more of its probability in a small region around a.

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 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—but in certain applications 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[23]

$$\mathcal{V}(X) \equiv \frac{1}{n}\mathcal{E}\{[\sqrt{n}(X_n - \mu_n)]^2\} \tag{D.7}$$

---

[23]It is generally preferable to define asymptotic expectation and variance in terms of the asymptotic

D.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 $n(\epsilon)$ such that $|P_n(x) - P(x)| < \epsilon$ for all $n > n(\epsilon)$ 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.

The results of this section extend straightforwardly to vectors and matrices: We say that $\operatorname{plim} \underset{(m \times 1)}{\mathbf{x}} = \underset{(m \times 1)}{\mathbf{a}}$ when $\operatorname{plim} X_i = a_i$ for $i = 1, 2, \ldots, m$. Likewise, $\operatorname{plim} \underset{(m \times p)}{\mathbf{X}} = \underset{(m \times p)}{\mathbf{A}}$ means that $\operatorname{plim} X_{ij} = a_{ij}$ 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

$$\mathcal{V}(\mathbf{x}) \equiv \frac{1}{n}\mathcal{E}\{[\sqrt{n}(\mathbf{x}_n - \boldsymbol{\mu}_n)][\sqrt{n}(\mathbf{x}_n - \boldsymbol{\mu}_n)]'\}$$

D.5 Properties of Estimators[24]

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.

D.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

$$S_*^2 \equiv \frac{\sum(X_i - \overline{X})^2}{n} \tag{D.8}$$

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 D.16.

---

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.

[24]Most of the material in this and the following section can be traced to a remarkable, seminal paper on estimation by Fisher (1922).
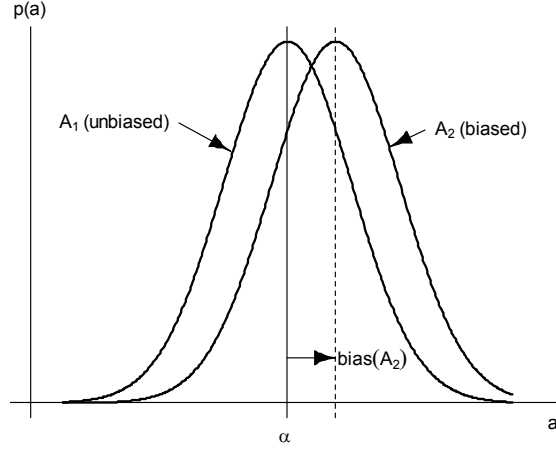
Figure D.16: The estimator A_1 is an unbiased estimator of α because $E(A_1) = \alpha$; the estimator A_2 has a positive bias, because $E(A_2) > \alpha$.

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 $n \to \infty$.

D.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: $\mathrm{MSE}(A) \equiv E[(A-\alpha)^2]$. 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,

$$
\begin{aligned}
\mathrm{MSE}(A) &= E[(A - \alpha)^2] \\
&= E\{[A - E(A) + E(A) - \alpha]^2\} \\
&= E\{[A - E(A)]^2\} + [E(A) - \alpha]^2 \\
&\quad + 2[E(A) - E(A)][E(A) - \alpha] \\
&= V(A) + [\mathrm{bias}(A)]^2 + 0
\end{aligned}
$$

The efficiency of an estimator increases, therefore, as its sampling variance and bias
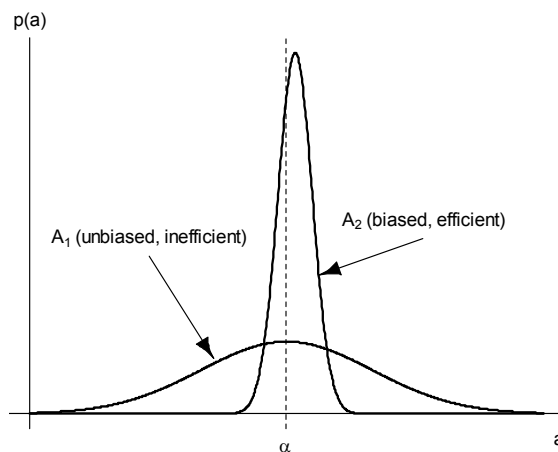
Figure D.17: Relative efficiency of estimators: Even though it is biased, A_2 is a more efficient estimator of α than the unbiased estimator A_1, because the smaller variance of A_2 more than compensates for its small bias.

decline. In comparing two estimators, an advantage in sampling variance can more than offset a disadvantage due to bias, as illustrated in Figure D.17.

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.

D.5.3   Consistency*

An estimator A of the parameter α is *consistent* if $\operatorname{plim} A = \alpha$. A sufficient (but not necessary[25]) 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 D.15 (page 88) illustrates consistency, if we construe X as an estimator of a. The estimator S_*^2 given in Equation D.8 (on page 89) is a consistent estimator of the population variance σ^2 even though it is biased in finite samples.

D.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

---

[25]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.

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 $1 - \pi$.[26] 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 1's and $n - s$ 0's, each with probability $1/\binom{n}{s}$.[27] 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.

D.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.

D.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:

$$
\begin{aligned}
\Pr(\text{data}|\text{parameter}) &= \Pr(HHTHHHTTHH|\pi) \\
&= \pi\pi(1 - \pi)\pi\pi\pi(1 - \pi)(1 - \pi)\pi\pi \\
&= \pi^7(1 - \pi)^3
\end{aligned}
$$

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

---

[26] 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.

[27] The random variable S has a binomial distribution: See Section D.2.1.

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*:

$$L(\text{parameter}|\text{data}) = L(\pi|HHTHHHTTHH)$$
$$= \pi^7(1-\pi)^3$$

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 D.18. 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.

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 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 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

$$\log_e L(\pi) = x\log_e \pi + (n-x)\log_e(1-\pi) \tag{D.9}$$

---

[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 $(2^{10} = 1024)$ of possible samples.
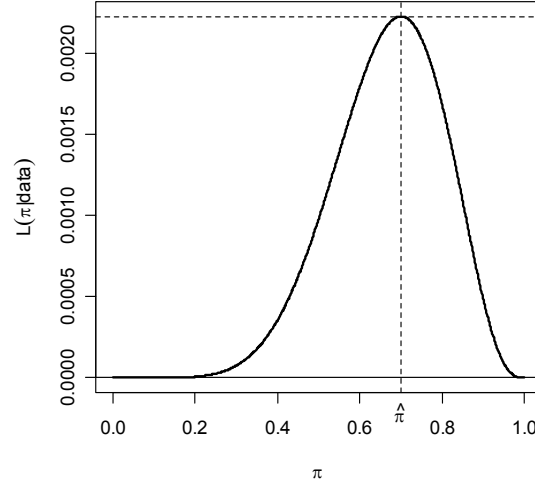
Figure D.18: The likelihood function $L(\pi|HHTHHHTTHH) = \pi^7(1-\pi)^3$.

Differentiating $\log_e L(\pi)$ with respect to π produces

$$\begin{aligned}\frac{d\log_e L(\pi)}{d\pi} &= \frac{x}{\pi} + (n-x)\frac{1}{1-\pi}(-1)\\ &= \frac{x}{\pi} - \frac{n-x}{1-\pi}\end{aligned}$$

Setting the derivative to 0 and solving for π produces the MLE which, as before, is the sample proportion x/n. The maximum-likelihood *estimator* is $\widehat{\pi} = X/n$. To avoid this slightly awkward substitution of estimator for estimate in the last step, we usually replace x by X in the log likelihood function (Equation D.9).

D.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 appendix):

- 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.

- The asymptotic sampling variance of the MLE $\widehat{\alpha}$ of a parameter α can be obtained from the second derivative of the log likelihood:

$$\mathcal{V}(\widehat{\alpha}) = \frac{1}{-E\left[\dfrac{d^2 \log_e L(\alpha)}{d\alpha^2}\right]} \tag{D.10}$$

The denominator of $\mathcal{V}(\widehat{\alpha})$ is called the *expected or Fisher information*,

$$\mathcal{I}(\alpha) \equiv -E\left[\frac{d^2 \log_e L(\alpha)}{d\alpha^2}\right]$$

In practice, we substitute the MLE $\widehat{\alpha}$ into Equation D.10 to obtain an *estimate* of the asymptotic sampling variance, $\widehat{\mathcal{V}}(\widehat{\alpha})$.[29]

- $L(\widehat{\alpha})$ 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*

$$G^2 \equiv -2\log_e \frac{L(\alpha)}{L(\widehat{\alpha})} = 2[\log_e L(\widehat{\alpha}) - \log_e L(\alpha)]$$

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).

- 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.

Establishing these results is well beyond the scope of this appendix, 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.

D.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*[30]—for testing the statistical hypothesis H_0: $\alpha = \alpha_0$. The Wald and likelihood-ratio tests can be "turned around" to produce confidence intervals for α.

---

[29]It is also possible, and sometimes computationally advantageous, to base an estimate of the variance of the MLE $\widehat{\alpha}$ on the *observed information*,

$$\mathcal{I}_O(\widehat{\alpha}) \equiv \frac{d^2 \log_e L(\widehat{\alpha})}{d\widehat{\alpha}^2}$$

[30]The score test is sometimes called the *Lagrange-multiplier test*. (Lagrange multipliers are described in Section C.5.2.)
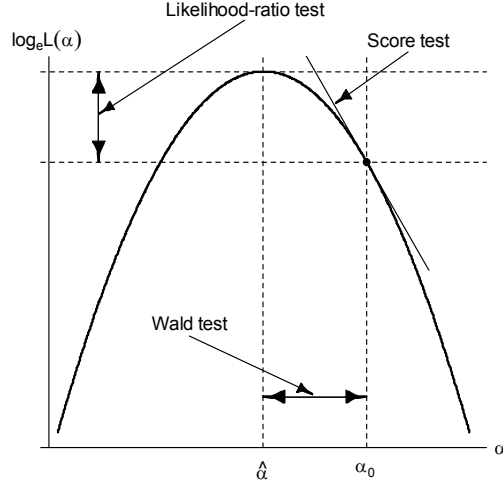
Figure D.19: Tests of the hypothesis H_0: $\alpha = \alpha_0$: The likelihood-ratio test compares $\log_e L(\widehat{\alpha})$ to $\log_e L(\alpha_0)$; the Wald test compares $\widehat{\alpha}$ to α_0; and the score test examines the slope of $\log_e L(\alpha)$ at $\alpha = \alpha_0$.

- *Wald test.* Relying on the asymptotic[31] 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

$$G_0^2 \equiv -2 \log_e \frac{L(\alpha_0)}{L(\widehat{\alpha})} = 2[\log_e L(\widehat{\alpha}) - \log_e L(\alpha_0)]$$

  is asymptotically distributed as χ_1^2 under H_0.

- *Score test.* The "score" $S(\alpha)$ is the slope of the log likelihood at a particular value of α.[32] 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 S_0 depends only on the null value α_0, which is

---

[31]Asymptotic results apply approximately, with the approximation growing more accurate as the sample size n gets larger.

[32]*That is, $S(\alpha) \equiv d \log_e L(\alpha)/d\alpha$.

specified in H_0). In most instances, however, the LR test is more reliable than the Wald and score tests in smaller samples.

Figure D.19 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 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 D.9 on page 93) 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.[33] 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

$$\pi = .7 \pm 1.96 \times \sqrt{0.0210} = .7 \pm .290$$

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,

$$\log_e L(\widehat{\pi}) = \log_e L(.7) = 7 \log_e .7 + 3 \log_e .3 = -6.1086$$
$$\log_e L(\pi_0) = \log_e L(.5) = 7 \log_e .5 + 3 \log_e .5 = -6.9315$$

---

[33]In this case, the asymptotic variance coincides with the exact, finite-sample variance of $\widehat{\pi}$.

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 $S(\pi_0) = 7/.5 - 3/.5 = 8$. Evaluating the expected information at π_0 produces $\mathcal{I}(\pi_0) = \mathcal{I}(.5) = 10/(.5 \times .5) = 40$. 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),

$$p(x) = \binom{10}{x} .5^x .5^{10-x} = \binom{10}{x} .5^{10}$$

yields a two-tail p-value of .3438 [corresponding to $\Pr(X \le 3 \text{ or } X \ge 7)$].[34] We must be careful in applying asymptotic results to small samples.

D.6.4   Several Parameters*

The maximum-likelihood method can be generalized to simultaneous estimation of several parameters. Let $p\left(\underset{(n \times m)}{\mathbf{X}} \mid \underset{(k \times 1)}{\boldsymbol{\alpha}}\right)$ represent the probability or probability density for n possibly multivariate observations $\mathbf{X}$ ($m \ge 1$) which depend on k independent parameters $\boldsymbol{\alpha}$.[35] The likelihood $L(\boldsymbol{\alpha}) \equiv L(\boldsymbol{\alpha}|\mathbf{X})$ 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

$$\underset{(k \times k)}{\mathcal{V}(\widehat{\boldsymbol{\alpha}})} = \left\{ -E\left[\frac{\partial^2 \log_e L(\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}\,\partial\boldsymbol{\alpha}'} \right] \right\}^{-1} \tag{D.11}$$

The matrix in braces in Equation D.11 is called the *expected* or *Fisher information matrix*, $\mathcal{I}(\boldsymbol{\alpha})$ (not to be confused with the identity matrix $\mathbf{I}$).[36] Moreover, if $\boldsymbol{\beta} = f(\boldsymbol{a})$, then the MLE of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}} = f(\widehat{\boldsymbol{a}})$. Notice how the formulas for several parameters closely parallel those for one parameter.

---

[34] See Section D.2.1 for a discussion of the binomial distribution.

[35] 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.

[36] As before, it is also possible to work with the *observed* information at the MLE $\widehat{\boldsymbol{\alpha}}$.

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:

$$G_0^2 \equiv -2 \log_e \left[\frac{L(\boldsymbol{\alpha}_0)}{L(\widehat{\boldsymbol{\alpha}})} \right]$$

All three test statistics are asymptotically distributed as χ_k^2 under H_0.

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,

$$G_0^2 \equiv -2 \log_e \left[\frac{L(\widehat{\boldsymbol{\alpha}}_0)}{L(\widehat{\boldsymbol{\alpha}})} \right]$$

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

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(X_i - \mu)^2}{2\sigma^2} \right]$$

$$= (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n}(X_i - \mu)^2 \right]$$

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

$$\frac{\partial \log_e L(\mu, \sigma^2)}{\partial\mu} = \frac{1}{\sigma^2}\sum(X_i - \mu)$$

$$\frac{\log_e L(\mu, \sigma^2)}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum(X_i - \mu)^2$$

Setting the partial derivatives to 0 and solving simultaneously for the maximum-likelihood estimators of μ and σ^2 produces

$$\widehat{\mu} = \frac{\sum X_i}{n} = \overline{X}$$

$$\widehat{\sigma}^2 = \frac{\sum(X_i - \overline{X})^2}{n}$$

The matrix of second partial derivatives of the log likelihood is

$$
\begin{bmatrix}
\dfrac{\partial^2 \log_e L}{\partial \mu^2} & \dfrac{\partial^2 \log_e L}{\partial \mu \, \partial \sigma^2} \\[2ex]
\dfrac{\partial^2 \log_e L}{\partial \sigma^2 \, \partial \mu} & \dfrac{\partial^2 \log_e L}{\partial (\sigma^2)^2}
\end{bmatrix}
=
\begin{bmatrix}
-\dfrac{n}{\sigma^2} & -\dfrac{1}{\sigma^4}\sum (X_i - \mu) \\[2ex]
-\dfrac{1}{\sigma^4}\sum (X_i - \mu) & \dfrac{n}{2\sigma^4} - \dfrac{1}{\sigma^6}\sum (X_i - \mu)^2
\end{bmatrix}
$$

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:

$$
-\mathcal{I}(\mu, \sigma^2) =
\begin{bmatrix}
-\dfrac{n}{\sigma^2} & 0 \\[2ex]
0 & -\dfrac{n}{2\sigma^4}
\end{bmatrix}
$$

The asymptotic variance-covariance matrix of the maximum-likelihood estimators is, as usual, the inverse of the information matrix:

$$
\mathcal{V}(\widehat{\mu}, \widehat{\sigma}^2) = [\mathcal{I}(\mu, \sigma^2)]^{-1} =
\begin{bmatrix}
\dfrac{\sigma^2}{n} & 0 \\[2ex]
0 & \dfrac{2\sigma^4}{n}
\end{bmatrix}
$$

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 D.8 on page 89).

D.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 **??**) to $f(\widehat{\alpha})$ around the true value of the parameter α:

$$
\widehat{\beta} = f(\widehat{\alpha}) \approx f(\alpha) + f'(\alpha)(\widehat{\alpha} - \alpha) \tag{D.12}
$$

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 D.12, $f(\alpha)$, 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

$$
\mathcal{V}(\widehat{\beta}) = [f'(\alpha)]^2 \, \mathcal{V}(\widehat{\alpha})
$$

where $\mathcal{V}(\widehat{\alpha})$ is the asymptotic variance of $\widehat{\alpha}$. In practice, we substitute the MLE $\widehat{\alpha}$ for α to obtain the *estimated* asymptotic variance of $\widehat{\beta}$:

$$
\widehat{\mathcal{V}}(\widehat{\beta}) = [f'(\widehat{\alpha})]^2 \, \mathcal{V}(\widehat{\alpha})
$$

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 $\mathcal{V}(\widehat{\pi}) = \pi(1 - \pi)/n$, 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

$$\begin{aligned}
\mathcal{V}(\widehat{\Lambda}) &= [f'(\pi)]^2 \, \mathcal{V}(\widehat{\pi}) \\
&= \left[\frac{1}{\pi(1 - \pi)}\right]^2 \frac{\pi(1 - \pi)}{n} \\
&= \frac{1}{n\pi(1 - \pi)}
\end{aligned}$$

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

$$\mathcal{V}(\widehat{\beta}) = [\mathbf{g}(\boldsymbol{\alpha})]' \, \mathcal{V}(\widehat{\boldsymbol{\alpha}})\mathbf{g}(\boldsymbol{\alpha}) = \sum_{i=1}^{k}\sum_{j=1}^{k} v_{ij} \times \frac{\partial\widehat{\beta}}{\partial\alpha_i} \times \frac{\partial\widehat{\beta}}{\partial\alpha_j}$$

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

$$\widehat{\mathcal{V}}(\widehat{\beta}) = [\mathbf{g}(\widehat{\boldsymbol{\alpha}})]' \, \mathcal{V}(\widehat{\boldsymbol{\alpha}})\mathbf{g}(\widehat{\boldsymbol{\alpha}})$$

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.

D.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.

D.8 Introduction to Bayesian Inference

This section introduces Bayesian statistics, an alternative approach to statistical inference. The treatment here is very brief because Bayesian methods are used at only two points in the text: multiple imputation of missing data (in Chapter 20), and model selection and Bayesian model averaging (in Chapter 22).

D.8.1 Bayes' Theorem

Recall (from Section D.1.1) the definition of *conditional probability*: The probability of an event A given that another event B is known to have occurred is

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} \tag{D.13}$$

Likewise, the conditional probability of B given A is

$$\Pr(B|A) = \frac{\Pr(A \cap B)}{\Pr(A)} \tag{D.14}$$

Solving Equation D.14 for the *joint probability* of A and B produces

$$\Pr(A \cap B) = \Pr(B|A)\Pr(A)$$

and substituting this result into Equation D.13 yields *Bayes' Theorem*:[37]

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)} \tag{D.15}$$

Bayesian statistical inference is based on the following interpretation of Equation D.15: 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[38]

$$\Pr(B) = \Pr(B|A)\Pr(A) + \Pr(B|\overline{A})\Pr(\overline{A})$$

Then $\Pr(A|B)$, given by Equation D.15 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.[39]

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, 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.[40] 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

$$\Pr(B|A) = .3^7(1-.3)^3 = .0000750$$
$$\Pr(B|\overline{A}) = .8^7(1-.8)^3 = .0016777$$

---

[37] Bayes' theorem is named after its discoverer, the Reverend Thomas Bayes, an 18th Century English mathematician.

[38] This is an application of the *law of total probability*: Given an event B and a set of 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),

$$\Pr(B) = \sum_{i=1}^{k} \Pr(B|A_i)\Pr(A_i)$$

[39] The frequentist interpretation of probabilities is described in Section D.1.1. Bayes' 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.

[40] These are the data used in a preliminary example of maximum-likelihood estimation in Section D.6.

Notice that, as is typically the case, the likelihood of the observed data is small in any event, but the data are much more likely under $\overline{A}$ than under A.[41] Using Bayes' Theorem (Equation D.15), 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 $\Pr(H) = .3$.

D.8.2   Extending Bayes Theorem

Bayes' 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;[42] 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

$$\Pr(H_i|D) = \frac{\Pr(D|H_i)\Pr(H_i)}{\sum_{j=1}^{k} \Pr(D|H_j)\Pr(H_j)} \tag{D.16}$$

The denominator in Equation D.16 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' 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:

---

[41] The likelihood of these data for *any* value of $\Pr(H)$ between 0 and 1 was shown previously in Figure D.18 (page 94).

[42] To employ Bayesian inference, your prior beliefs must be consistent with probability theory, and so the prior probabilities must sum to 1.

- 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.

D.8.3   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 likelihood is

$$L(\pi) = \pi^h(1-\pi)^{10-h} \tag{D.17}$$

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 likelihood in Equation D.17 is the beta distribution[43]

$$p(\pi) = \frac{\pi^{a-1}(1-\pi)^{b-1}}{B(a,b)} \text{ for } 0 \le \pi \le 1 \text{ and } a,b \ge 1$$

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 $h+a-1 = 6+a$ and $10-h+b-1 = 2+b$. 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.[44] If you are uncomfortable with this restriction, then you could select smaller

---

[43]See Section D.3.8.
[44]See Figure D.13 on page 86.

values of a and b: In the extreme, $a = b = 1$, and all values of π are equally likely—a so-called *flat prior distribution*, reflecting complete ignorance about the value of π.[45]

Figure D.8.3 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 $\widehat{\pi} = .7$.[46] 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 $a + b - 2$ 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.[47] In the current example, if the coin is flipped n times, then the posterior distribution takes the form

$$p(\pi|D) \propto \pi^{h+a-1}(1 - \pi)^{n-h+b-1}$$

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.

D.8.4 Bayesian Interval Estimates

As in classical inference, it is desirable not only to provide a point estimate of a parameter but also to express uncertainty in the estimate. The posterior distribution of the parameter expresses 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-percent posterior interval. This difference reflects the Bayesian interpretation of a parameter as a random variable, with the posterior distribution expressing subjective uncertainty in 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 D.8.3.

---

[45] 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 $\equiv \log_e [\pi/(1 - \pi)]$; a flat prior for π is not flat for ω or for the logit.

[46] 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.

[47] 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.
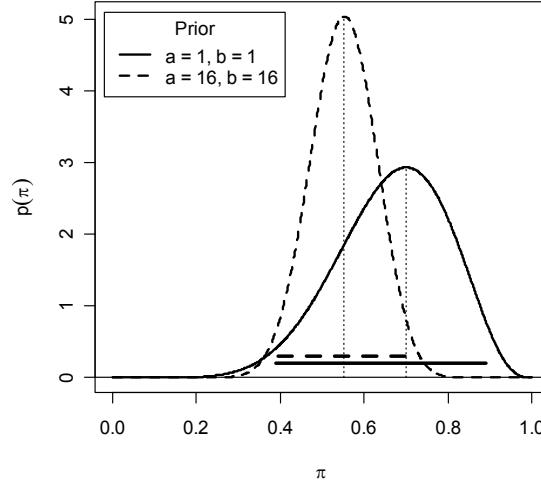
D.8.5   Bayesian Inference for Several Parameters

Bayesian inference extends straightforwardly to the simultaneous estimation of several parameters $\boldsymbol{\alpha} \equiv [\alpha_1, \alpha_2, \ldots, \alpha_p]'$. 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 focusses on the *marginal posterior distribution* of each parameter, $p(\alpha_i|D)$.

D.9   Recommended Reading

Almost any introductory text in mathematical statistics, and many econometric texts, cover the subject matter of this appendix more formally and in greater detail. Cox and Hinkley (1974) is a standard, if relatively difficult, treatment of most of the topics in this appendix. 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 un-starred parts of this appendix too terse. A good, relatively accessible discussion of asymptotic distribution theory appears in Theil (1971, Chapter 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 introduction to Bayesian methods.

References

Binmore, K. G. (1983). *Calculus*. Cambridge University Press, Cambridge.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman and Hall, London.

Davis, P. J. (1965). *Mathematics of Matrices: A First Book of Matrix Theory and Linear Algebra*. Blaisdell, New York.

Engle, R. F. (1984). Wald, likelihood ratio, and Lagrange multiplier tests in econometrics. In Griliches, Z. and Intriligator, M. D., editors, *Handbook of Econometrics*, volume II, pages 775–879. North-Holland, Amsterdam.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, A*, 222:309–368.

Graybill, F. A. (1983). *Introduction to Matrices With Applications in Statistics*. Wadsworth, Belmont CA, second edition.

Green, P. E. and Carroll, J. D. (1976). *Mathematical Tools for Applied Multivariate Analysis*. Academic Press, New York.

Healy, M. J. R. (1986). *Matrices for Statistics*. Clarendon Press, Oxford.

Johnston, J. (1972). *Econometric Methods*. McGraw-Hill, New York, second edition.

Kennedy, W. J., Jr., and Gentle, J. E. (1980). *Statistical Computing*. Dekker, New York.

Lancaster, T. (2004). *An Introduction to Modern Bayesian Econometrics*. Blackwell, Oxford.

McCallum, B. T. (1973). A note concerning asymptotic covariance expressions. *Econometrica*, 41:581–583.

Monahan, J. F. (2001). *Numerical Methods of Statistics*. Cambridge University Press, Cambridge.

Namboodiri, K. (1984). *Matrix Algebra: An Introduction*. Sage, Beverly Hills, CA.

Rao, C. R. and Mitra, S. K. (1971). *Generalized Inverse of Matrices and Its Applications*. Wiley, New York.

Searle, S. R. (1982). *Matrix Algebra Useful for Statistics*. Wiley, New York.

Theil, H. (1971). *Principles of Econometrics*. Wiley, New York.

Thompson, S. P. and Gardner, M. (1998). *Calculus Made Easy*. St. Martin's, New York.

Wonnacott, R. J. and Wonnacott, T. H. (1979). *Econometrics*. Wiley, New York, second edition.

Wonnacott, T. H. and Wonnacott, R. J. (1990). *Introductory Statistics*. Wiley, New York, fifth edition.

Zellner, A. (1983). Statistical theory and econometrics. In Griliches, Z. and Intriligator, M. D., editors, *Handbook of Econometrics, Volume 1*, pages 67–178. North-Holland, Amsterdam.