Daniel L. Hartl
Harvard University

Andrew G. Clark
Pennsylvania State University

Principles of
Population Genetics

THIRD EDITION

Sinauer Associates, Inc. Publishers
Sunderland, Massachusetts
THE COVER
This image represents data from the first study of nucleotide sequence variation in a natural population, conducted by Martin Kreitman (1983, Nature 304:412-417). Each of the 11 vertical bands represents one allele of the gene alcohol dehydrogenase taken from a global distribution of the fruit fly Drosophila melanogaster. The colors correspond to the bases at each site: adenine = green, cytosine = yellow, guanine = blue, and thymine = red. Note that at each position only two different bases were observed. Blocks of sites in strong linkage disequilibrium can also be seen as repeated patterns of color. The sequences are oriented with the 5' end at the top, and the nucleotide corresponding to the fast/slow difference in the ADH protein is the twelfth one from the bottom.

To Barbara and Christine

PRINCIPLES OF POPULATION GENETICS, Third Edition
Copyright © 1997 by Sinauer Associates, Inc. All rights reserved.
This book may not be reproduced in whole or in part without permission from the publisher.

For information or to order, address
Sinauer Associates, Inc. PO Box 107
23 Plumtree Road, Sunderland, MA, 01375 U.S.A.
FAX 413-549-1118
Internet: publish@sinauer.com; http://www.sinauer.com

Library of Congress Cataloging-in-Publication Data
Hartl, Daniel L.
Cm
Includes bibliographical references and index
ISBN 0-87893-306-9 (hardcover)
1. Population genetics. 2. Quantitative genetics. 3. Population genetics—Problems, exercises, etc. I. Clark, Andrew G., 1954—. II. Title
QH455.H36 1997
576.58—dc21
97-34505
CIP

Printed in Canada
5 4 3 2 1
## Table of Contents

1. **GENETIC AND STATISTICAL BACKGROUND**
   - Gene Expression and Gene Interaction 2
     - Gene Expression 3
     - The Genetic Code 6
     - Alleles 8
   - Genotype and Phenotype 8
   - Segregation and Gene Interaction 11
   - Segregation and Recombination 12

2. **Probability in Population Genetics**
   - The Addition Rule 16
   - The Multiplication Rule 16
   - Repeated Trials 17

3. **Phenotypic Diversity and Genetic Variation**
   - Allele Frequencies in Populations 20
   - Parameters and Estimates 22
   - The Standard Error of an Estimate 22

4. **Models in Population Genetics**
   - Exponential Population Growth 27
   - Logistic Population Growth 31

5. **Random Mating**
   - Nonoverlapping Generations 73
   - The Hardy-Weinberg Principle 74

6. **Implications of the Hardy-Weinberg Principle**
   - Random Mating of Genotypes versus Random Union of Genotypes 76
   - Frequency of Heterozygotes 87
   - Special Cases of Random Mating 88
     - Three or More Alleles 88
     - X-Linked Genes 92

7. **Linkage and Linkage Disequilibrium**
   - Central Limit Theorem 41
   - Discrete Mendelian Variation 43
   - Experimental Methods for Detecting Genetic Variation 44
     - Protein Electrophoresis 45
     - The Southern Blot Procedure 48
     - The Polymerase Chain Reaction 51
   - Polymorphism and Heterozygosity 53
     - Allozyme Polymorphisms 54
     - How Representative Are Allozymes? 56
     - Polymorphisms in DNA Sequences 57
     - Nucleotide Polymorphism and Nucleotide Diversity 57
     - Uses of Genetic Polymorphisms 62
   - Multiple-locus Inheritance 64

8. **Organizational Genetic Variation**
   - Random Mating 72
     - Nonoverlapping Generations 73
     - The Hardy-Weinberg Principle 74
     - Random Mating of Genotypes versus Random Union of Genes 76
   - Implications of the Hardy-Weinberg Principle 79
     - The Hardy-Weinberg Principle in Operation 80
     - Complications of Dominance 84
     - Frequency of Heterozygotes 87
   - Special Cases of Random Mating 88
     - Three or More Alleles 88
     - X-Linked Genes 92
   - Linkage and Linkage Disequilibrium 95

9. **Population Substructure**
   - Hierarchical Population Structure 111
     - Reduction in Heterozygosity 112
     - Wright's F Statistics 117
   - Genetic Diversity among Subpopulations 120
   - Isolate Breaking: The Wahlund Principle 122
     - Wahlund's Principle and the Fixation Index 125
     - Genetic Frequencies in Subdivided Populations 127

10. **Population Genetics in DNA Typing**
    - Polymorphisms Based on a Variable Number of Tandem Repeats (VNTR) 129
    - Match Probabilities with Hardy-Weinberg Equilibrium and Linkage Equilibrium 132
    - Effects of Population Substructure 132

11. **Inbreeding**
    - Genetic Frequencies with Inbreeding 135
    - Relation between the Inbreeding Coefficient and the F Statistics 139
    - The Inbreeding Coefficient as a Probability 141
    - Genetic Effects of Inbreeding 145
    - Calculation of the Inbreeding Coefficient from Pedigrees 149
    - Regular Systems of Mating 153
    - Assortative Mating 155

12. **DARWINIAN SELECTION**
    - Selection in Haploid Organisms 212
      - Discrete Generations 212
      - Continuous Time 216
      - Change in Allele Frequency in Haploids 217

13. **SOURCES OF VARIATION**
    - Mutation 163
      - Irreversible Mutation 164
      - Reversible Mutation 168
      - Probability of Fission of a New Neutral Mutation 170
      - The Infinite-Alleles Model 174
      - Neutral Mutations 177
    - Linkage and Recombination 180
      - Pressured Evolutionary Benefit of Recombination 181
      - Recombination and Polymorphism 181
      - Precise Recombination in Bacteria 186
      - Absence of Recombination in Annual Mateless Cytogenetic DNA 187

14. **Migration**
    - One-Way Migration 189
    - The Island Model of Migration 192
    - How Migration Limits Genetic Drift 194
    - Estimates of Migration Rates 196
    - Patterns of Migration 196

15. **Transposable Elements**
    - Factors Controlling the Population Dynamics of Transposable Elements 200
    - Insertion Sequences and Composite Transposons in Bacteria 200
    - Transposable Elements in Eukaryotes 204
    - Horizontal Transmission of Transposable Elements 204

16. **SUMMARY**
    - Summary 206
    - Problems 208

17. **REFERENCES**
Table of Contents

Darwinian Fitness and Mullerian Fitness 218
Selection in Diploid Organisms 218
Change in Allele Frequency in Diploids 219
Time Required for a Given Change in Allele Frequency 222
Application to the Evolution of Instericidal Resistance 226
Equilibria with Selection 227
Overdominance 228
Local Stability 232
Heterozgyote Incompatibility 234
The Adaptive Topography and the Role of Random Genetic Drift 236
Mutation-Selection Balance 236
Equilibrium Allele Frequencies 237
The Haldane-Muller Principle 239
More Complex Types of Selection 240
Frequency-Dependent Selection 240
Density-Dependent Selection 240
Facilitative Selection 241
Age-Structured Populations 242
Heterogeneous Environments and Clines 242
Diversifying Selection 244
Differential Selection in the Sexes 246
X-Linked Genes 246
Genetic Selection 246
Genetic Drive 247
Multiple Alleles 250
Multiple Loci and Gene Interaction: Epistasis 252
Sexual Selection 255
Kin Selection 256
Interdeme Selection and the Shifting Balance Theory 259
Summary 262
Problems 264

7 RANDOM GENETIC DRIFT 267
Random Genetic Drift and Binomial Sampling 267
The Wright-Fisher Model of Random Genetic Drift 274
The Diffusion Approximation 277
Absorption Time and Time to Fixation 282
Parallels between Random Drift and Inbreeding 283
Effective Population Size 289
Fluctuation in Population Size 290
Unequal Sex Ratio, Sex Chromosomes, Organellar Genes 292
Balance between Mutation and Drift 294
Infinite Alleles Model 294
The Ewens Sampling Formula 296
The Ewens-Watson Test 298
Infinite-Sites Model 300
Gene Trees and the Coalescent 304
Coalescent Models with Mutation 308
Summary 310
Problems 312

8 MOLECULAR POPULATION GENETICS 315
The Neutral Theory and Molecular Evolution 315
Theoretical Principles of the Neutral Theory 316
Estimating Rates of Molecular Sequence Divergence 320
Rates of Amino Acid Replacement 320
Rates of Nucleotide Substitution 324
Other Measures of Molecular Divergence 327
The Molecular Clock 328
Variation across Genes in the Rate of the Molecular Clock 331
Variation across Lineages in Clock Rate 333
The Generation-Time Effect 336
Does the Constancy of Substitution Rates Prove the Neutral Theory? 337
Patterns of Nucleotide and Amino Acid Substitution 338
Calculating Synonymous and Nonsynonymous Substitution Rates 338
Within-Species Polymorphism 345
Implications of Codon Bias 348
Polymorphism and Divergence in Nucleotide Sequence Data 349
Impact of Local Recombination Rates 353
Gene Genealogies 354
Hypothesis Testing Using Trees 356
Inferences about Migration Based on Gene Trees 360
Mitochondrial and Chloroplast DNA Evolution 361
Chloroplast DNA and Organelle Transmission in Plants 365
Maintenance of Variation in Organelle Genomes 366
Evidence for Selection in mtDNA 367
Molecular Phylogenetics 368
Algorithms for Phylogenetic Tree Reconstruction 368
Distance Methods versus Parsimony 372
Bootstrapping and Statistical Confidence in a Tree 372
Shared Polymorphism 373
Interspecific Genotypes 374
Multigene Families 374
Causes of Convergent Evolution 375
Multigene Family Evolution through a Drift and Death Process 378
Structural RNA Genes and Compensatory Substitutions 382
Multigene Superfamilies 383
Dispersed Highly Repeated DNA Sequences 385
Summary 390
Problems 392

9 QUANTITATIVE GENETICS 397
Types of Quantitative Traits 398
Resemblance between Relatives and the Concept of Heritability 400
Artificial Selection and Realized Heritability 406
Prediction Equation for Individual Selection 410
Selection Limits 411
Genetic Models for Quantitative Traits 414
Change in Gene Frequency 421
Genetic Model for the Change in Mean Phenotype 423
Components of Phenotypic Variance 424
Genetic and Environmental Sources of Variation 425
Components of Genetic Variation 430
Covariance between Relatives 434
Twin Studies and Inferences of Heritability in Humans 440
Experimental Assessment of Genetic Variance Components 442
Indirect Estimation of the Number of Genes Affecting a Quantitative Character 445
Norm of Reaction and Phenotypic Plasticity 448
Threshold Traits and the Genetics of Liability 452
Correlated Response and Genetic Correlation 453
Inference of Selection from Phenotypic Data 458
Evolution of Quantitative Traits 460
Preface

Thanks in part to the power of molecular methods, population genetics has been reinvigorated. As some genome projects are approaching closure and methods of "functional genomics" are scaling up to identify the roles of novel genes, inevitably increasing attention is being paid to the significance of genetic variation in populations. Nowhere is this more evident than in medical genetics. Within a decade we can expect that all major single-gene inherited disorders will be identified, genetically mapped, cloned, and characterized at a fine molecular level. Health professionals realize that this impressive feat will have an impact only on a small minority of individuals. Most of the genetic variation in disease risk is multifactorial, which means that the risk is determined by multiple genetic and environmental factors acting together. Killer diseases such as familial forms of cancer, diabetes, and cardiovascular disease fall into this category. The fact that these diseases aggregate in families implies that there is probably a genetic component, but the genetic component may differ from one family or ethnic group to another.

Prompted by the high incidence of multifactorial diseases as a group, the medical community has become acutely aware of the need to understand the basic structure of genetic variation in populations in order to determine what aspects of the variation cause disease.

The exciting practical applications of population genetics to the analysis of multifactorial diseases have received great attention, but the scope of population genetics actually is much broader. Population genetics provides the genetic underpinning for all of evolutionary biology. By "evolution" we mean descent with modification. Species undergo progressive genetic modification as they adapt to their environments, and new species arise as a by-product of this process. The intellectual excitement of biological evolution arises from the fact that it addresses the fundamental questions, "What are we?" and "Where did we come from?"

Patterns of evolutionary history are recorded in DNA sequences, and the application of population genetics to interpreting DNA sequences is revealing many secrets about the evolutionary past, including the history of our own species. But population genetics embraces much more than the analysis of evolutionary relationships. It is particularly concerned with the processes and mechanisms by which evolutionary changes are made. The field is inherently multidisciplinary, cutting across molecular biology, genetics, ecology, evolutionary biology, systematics, natural history, plant breeding, animal breeding, conservation and wildlife management, human genetics, sociology, anthropology, mathematics, and statistics.

Students taking population genetics are usually expected to have completed, or to be taking concurrently, a course in differential calculus. While this book assumes a familiarity with the elementary notation for differentials and integrals, it does not require...
great mathematical proficiency. We have kept the mathematics to a minimum. On the other hand, some of the most important ideas in population genetics require quite advanced mathematics. Rather than ignore these approaches, we have made a concerted effort to present these models in such a way that the assumptions can be understood and the main results appreciated without much mathematics. References are provided for the interested reader to learn more about the details.

Several important changes distinguish the third edition of *Principles* from the second edition. The level of the treatment is more tailored to the needs of a one-semester or one-quarter course, with the intended audience being third- and fourth-year undergraduates as well as beginning graduate students. Population genetics is not only an experimental science but also a theoretical one. Special care has been taken to explain the biological motivation behind the theoretical models so that the models do not simply materialize out of thin air, and to explain in plain English the implications of the results. Many concepts are illustrated by numerical examples, using actual data wherever possible. Special topics and examples are often set off from the text as boxed problems whose solutions are explained step by step. Every chapter ends with about 20 problems, graded in difficulty, and solutions worked in full appear at the end of the text.

This edition of *Principles* is organized into nine chapters that gradually build concepts from measuring variation and the various forces that influence genetic variation through a sequential progression to concepts of molecular population genetics and quantitative genetics. The first chapter provides a background in basic genetic and statistical principles. We discuss the fundamental concepts of allelism, dominance, segregating, recombination, and population frequencies. The force of model building and testing in population genetics is emphasized. Chapter 2 introduces the student to the primary data of population genetics, namely, the many levels of genetic variation. Chapter 3 is concerned with the organization of genetic variation into genotypes in populations. Here the Hardy-Weinberg principle gets very thorough coverage, including the cases of X-linkage and multiple alleles. Chapter 4 widens the perspective and considers the organization of genetic variation among spatially structured populations. Population substructure is measured by Wright's *F* statistic, and is presented in a way that conveys their biological meaning. The Wahlund principle and inbreeding are also covered in Chapter 4.

The goal of population genetics is to understand the forces that have an impact on levels of genetic variation. The forces of mutation, recombination, and migration are outlined in Chapter 5. Darwinian selection is the topic of Chapter 6, including both the theoretical foundations and empirical observations of the dynamics of gene-frequency change under the action of selection. Haploid and diploid cases are developed, as are the concepts of equilibrium, stability, and context dependence. After classical models of mutation-selection balance are developed, a series of more complex scenarios of natural selection are presented.

Chapter 7 deals with random genetic drift. In the absence of other forces, allele and genotype frequencies change as a result of random sampling from one generation to another. The Wright-Fisher model and diffusion approximations are presented in such a way that the student gains an appreciation for the importance of random genetic drift. The process of the coalescence of genealogies is an important innovation in theoretical population genetics, and some of the basic concepts of coalescence are presented in Chapter 7.

In Chapter 8 we cover the rapidly expanding data on molecular evolutionary genetics. The unifying theme in the study of molecular evolution is Kimura's neutral theory, and a close examination is made of the correspondence between the data and theory. This is a field in which advances in our empirical database and statistical tools for quantifying and manipulating the data are growing at a dizzying pace. Our goal is to give the student a firm grasp of the fundamentals, and a deep enough understanding of the principles to identify important gaps in our knowledge. One intriguing aspect of molecular evolutionary genetics is the discovery of new phenomena and forces taking place at the molecular level that go beyond the realm of classical population genetics. Multigene families and organelle genomes are described in some detail to illustrate these uniquely molecular phenomena.

Chapter 9 covers the problem of quantitative genetics from an evolutionary perspective. A compelling argument for using quantitative genetics for the study of evolution is that adaptive evolution takes place at the level of the phenotype, and quantitative genetics provides the tools for understanding transmission of phenotypic traits. Theoretical quantitative genetics is given special importance by the paradoxes it raises in contrasting evolution at the levels of the phenotype and of the DNA sequence. Our understanding of the correspondence between phenotypic and molecular differentiation is very incomplete, and our understanding of the correspondence between the rates of morphological and molecular evolution is even less well developed. As in the preceding chapters, we hope that the student is led with a feeling that there is plenty of room for imaginative work in this area. Population genetics is a field with a bright and expanding future.

ACKNOWLEDGMENTS

This book was greatly improved by the efforts of many people. The staff at Sinauer Associates did a splendid job assisting us with the revision. Nan Sinauer kept us on track, collecting and assembling dozens of computer files, revisions, FAQs, and email messages. Chris Small oversaw the page layout and managed the art program. Andy Sinauer played an essential role in having the book reviewed and in giving helpful advice as to level and length. We are grateful to Chip Aquadro, James Jacobson, Trudy Mackay, Roger Milkmak, Tim Prout, Glynnis Thomson, and Ken Weiss for their comments on the previous editions. Their insights greatly improved the presentation in this one.

Neither author could participate in writing a book such as this without a supportive, patient, sympathetic laboratory staff, able and willing to keep things running smoothly. The contents of this book are copyright © 2000 W. H. Freeman and Company. All rights reserved. No part of this book may be reproduced in any form without written permission from the publisher.

Several authors provided critical comments and suggestions: Sara Farace, Elena Lozovskaya, Dmitry Nurminsky, E. Fidelma Boyd, Allan Lohr, Jaceo Nagaraju, David Sullivan, Charles Hill, Dmitri Petrov, Mark Segal, Daniel De Aguilar, Carlos Bustamante, Jeffrey Townsend, Jorge Vieira, Christina Vieno, Isabel Beerman, Yusun Nam, Elizabeth Stover, and Susan Yukiens. In Andy Clark's laboratory, the acknowledgments include Michael Abraham, Joe Canala, Mancos Dernitzakis, Chris Fauco, Cristina Gonzalez, Ken Jordan, Angela Lamberti, Brian Lazzaro, Elf Masay, Harsh Spencer, Sarah Tishkoff, Bridget Todd, Carrie Tupper, and Lei Wang.
CHAPTER 1

Genetic and Statistical Background

**GENES** - **GENE EXPRESSION** - **PROBABILITY** - **ALLELE FREQUENCY ESTIMATES**

**STANDARD ERROR** - **MODELS** - **POPULATION GROWTH**

The science of **population genetics** deals with Mendel's laws and other genetic principles as they affect entire populations of organisms. The organisms may be human beings, animals, plants, or microbes. The populations may be natural, agricultural, or experimental. The environment may be city, farm, field, or forest. The habitat may be soil, water, or air. Because of its wide-ranging purview, population genetics cuts across many fields of modern biology. A working knowledge has become essential in genetics, evolutionary biology, systematics, plant breeding, animal breeding, ecology, natural history, forestry, horticulture, conservation, and wildlife management. A basic understanding of population genetics is also useful in medicine, law, biotechnology, molecular biology, cell biology, sociology, and anthropology.

Population genetics also includes the study of the various forces that result in evolutionary changes in species through time. By defining the framework within which evolution takes place, the principles of population genetics are basic to a broad evolutionary perspective on biology. From an experimental point of view, evolution provides a wealth of testable hypotheses for all other branches of biology. Many oddities in biology become comprehensible in the light of evolution: they result from shared ancestry among organisms, and they attest to the unity of life on earth.

Practical applications of population genetics are extensive. Many applications, particularly those relevant to human beings, also have important
implications in ethics and social policy. Among the applications of population genetics in medicine, agriculture, conservation, and research are:

- Genetic counseling of parents and other relatives of patients with hereditary diseases.
- Genetic mapping and identification of genes for disease susceptibility in human beings, including breast cancer, colon cancer, diabetes, schizophrenia, and so forth.
- Implications of population screening for carriers of disease genes, confidentiality of results, and maintenance of health insurability.
- Studies of the heritability of IQ score and its implications for affirmative action, welfare, and other social programs.
- Statistical interpretation of the significance of matching DNA types found between a suspect and a blood or semen sample from the scene of a crime.
- Design of studies to sample and preserve a record of genetic variation among human populations throughout the world.
- Improvement in the performance of domesticated animals and crop plants.
- Organization of mating programs for the preservation of endangered species in zoos and wildlife refuges.
- Sampling and preservation of germ plasms of potentially beneficial plants and animals that may soon vanish from the wild.
- Interpretation of differences in the nucleotide sequences of genes or amino acid sequences of proteins among members of the same or closely related species.

The genetic and statistical principles underlying population genetics are for the most part simple and straightforward, but it may be helpful to preface the discussion with a few key definitions and concepts.

**GENE EXPRESSION AND GENE INTERACTION**

Gene is a general term meaning, loosely, the physical entity transmitted from parent to offspring in reproduction that influences hereditary traits. Genes influence human traits such as hair color, eye color, skin color, height, weight, and various aspects of behavior—although most of these traits are also influenced more or less strongly by environment. Genes also determine the makeup of proteins such as hemoglobin, which carries oxygen in the red blood cells, or insulin, which is important in maintaining glucose balance in the blood. Genes can exist in different forms or states. For example, a gene for hemoglobin may exist in a normal form or in any one of a number of alternative forms that result in hemoglobin molecules that are more or less abnormal. These alternative forms of a gene are called alleles.

From a biochemical point of view, a gene corresponds to a region along a molecule of DNA (deoxyribonucleic acid). DNA is the genetic material. A molecule of DNA consists of two strands wound around each other in the form of a right-handed helix (the celebrated "double helix"). Each strand is a polymer of constituents called nucleotides, of which there are four, conventionally symbolized A, T, G, and C according to the nitrogen-rich base that each contains—either adenine (A), thymine (T), guanine (G), or cytosine (C). The paired strands are held together by weak chemical bonds (hydrogen bonds) that form between A and T at corresponding positions in opposite strands or between G and C at corresponding positions in opposite strands (Figure 1.1). Wherever one strand contains an A, the other across the way contains a T; and wherever one strand contains a G, the other across the way contains a C. Because of the pairing of complementary bases—A with T and G with C—a double-stranded DNA molecule contains an equal number of A and T nucleotides as well as an equal number of G and C nucleotides. DNA molecules can be very long. The DNA molecule in the bacterium *E. coli* is about 4.7 million base pairs, that in the largest chromosome in the fruit fly *Drosophila melanogaster* is about 65 million base pairs, and that in the largest human chromosome is about 230 million base pairs. Physical manipulation of such large molecules is impractical. In order to be studied, they must first be broken into smaller pieces.

**Gene Expression**

Most genes code for the polypeptide chains that constitute proteins. The code is the sequence of nucleotides along the DNA. In the decoding of the nucleotide sequence in DNA and also in the synthesis of proteins, several

![Figure 1.1](image-url) Genes are fundamental units of genetic information that correspond chemically to the sequence of nucleotides in a segment of DNA. A molecule of duplex DNA is composed of two intertwined strands, each of which consists of a long sequence of nucleotides. The strands are held together by pairing between the bases A and T in opposite strands and between the base G and C in opposite strands. The short diagonal lines indicate the paired bases. There are 10 base pairs per turn of the double helix. A typical gene consists of hundreds of thousands of nucleotides, only a few of which are shown here.
types of RNA (ribonucleic acid) are essential. RNA is also a polymer of nucleotides, each of which carries a base. Three of the bases in RNA (A, C, and G) are the same as those in DNA. The fourth (uracil (U)) is different. When an RNA strand pairs with a complementary strand of DNA, U in the RNA pairs with A in the DNA. Hence, the base-pairing role of U in RNA is the same as that of T in DNA.

The essentials of gene expression in the cells of higher organisms (eukaryotes) are outlined in Figure 1.2. The coding regions of the DNA in a gene, which code for amino acids, are often interrupted by one or more non-coding regions known as intervening sequences or introns. In the first step in gene expression (transcription), a molecule of RNA is produced that is complementary in base sequence to one of the strands of DNA (Figure 1.2A). Every gene includes a regulatory region (sometimes more than one) that determines when transcription takes place, the types of cells in which it takes place, and the strand that is to be transcribed. Because of the base pairing rules, a DNA sequence—say, 3'-ATCG-5'—results in a complementary RNA sequence—in this example, 5'-UAGC-3'. Note that the DNA and RNA strands each have a polarity or directionality. The terms 5' and 3' refer to the polarity of the strands. The 5' end typically terminates with a free phosphate group and the 3' end typically terminates with a free hydroxyl group (—OH).

When two strands of nucleic acid are paired, the polarity of each strand is opposite to that of the other. In the duplex DNA in Figure 1.2, for example, the left-to-right polarity of one strand is 5'-to-3', whereas the left-to-right polarity of the partner strand is 3'-to-5'. Similarly, in transcription, the template DNA strand has a left-to-right polarity of 5'-to-3', whereas the RNA transcript has the left-to-right polarity of 3'-to-5'. Because of the complementary base pairing between DNA and RNA nucleotides, the base-sequence code in DNA becomes converted into a base-sequence code in RNA. In transcription, the base sequence present in the introns is also faithfully copied into the base sequence of the RNA transcript.

The second step in gene expression in eukaryotes is RNA processing (Figure 1.2B). The beginning and end of the RNA transcript are chemically modified and the introns are removed by splicing (cutting and rearranging). RNA processing results in a molecule called messenger RNA (mRNA), in which the coding regions have been made contiguous. The regions in the original RNA transcript that are retained in the mature mRNA are called exons. The central part of the mRNA contains the spliced exons that code for the amino acid sequence of a polypeptide chain. The mRNA also includes exons upstream and downstream from the protein-coding region. The upstream region is the 5' untranslated region and the downstream region is the 3' untranslated region.

The final step in gene expression is translation, in which the mRNA molecule combines with ribosomes and other types of RNA molecules in the cytoplasm to produce the final polypeptide (Figure 1.2C). In the coding region of the mRNA, each adjacent group of three nucleotides constitutes a separate coding group or codon that specifies which amino acid is to be incorporated into the polypeptide chain. The ribosome moves along the mRNA in steps of three nucleotides (codon by codon). As each new codon comes into place, the correct amino acid is brought into line and attached to the end of the growing chain of amino acids. New amino acids are added to the growing chain until a codon specifying "stop" is encountered. At this point synthesis of the chain of amino acids is finished and the polypeptide is released from the ribosome.
In prokaryotes, which includes bacteria and other organisms lacking a nucleus, gene expression is essentially identical to that in eukaryotes except for the absence of RNA processing. Genes in prokaryotes do not contain introns and so splicing is unnecessary. In prokaryotes, the original RNA transcript is used immediately as mRNA and translated into a polypeptide. Because there is no separate nucleus, translation in prokaryotes often begins immediately when the 3' end of an RNA transcript comes off the DNA and even before transcription of the 3' end of the same molecule has been completed.

The central role of RNA in gene expression is one of the oddities of biology that makes sense in the light of evolution. That gene expression is configured around RNA is a legacy of the earliest forms of life when RNA molecules served both as carriers of genetic information and as catalytic molecules. The role of RNA as carrier of genetic information was gradually replaced by DNA, and the role of RNA as catalytic molecules was gradually replaced by proteins. At every step along the way, as the RNA world evolved into the DNA world, the role of RNA was indispensable in the processes of information transfer and protein synthesis, and so the RNA intermediates became locked in place.

The Genetic Code

The genetic code is the list of all codons showing which amino acid each codon specifies. Table 1.1 shows the standard genetic code used in nuclear genes in most organisms. A few organisms and some cellular organelles, such as mitochondria, use slightly altered codons. The codons in Table 1.1 are those found in the mRNA. The amino acids are given by three-letter abbreviations as well as by conventional single-letter abbreviations. Codon AUG is the start codon in polypeptide synthesis; it specifies methionine (Met) at the beginning of the polypeptide as well as at internal positions. Three codons are stops that result in termination of polypeptide synthesis: UAA, UAG, and UGA. The genetic code is redundant in that most amino acids are specified by more than one codon. Most of the redundancy is in the third codon position.

A code for an amino acid is twofold degenerate if either of two sequences specifies the same amino acid. Twofold degenerate codes have the pattern -Y or -R, where Y stands for any pyrimidine base (either U or C) and R stands for any purine base (either A or G). For example, CAU and CAC both code for histidine (His), fitting the pattern CAY, and CAA and CAG both code for glutamine (Gln), fitting the pattern CAR. A code for an amino acid is fourfold degenerate if any of four sequences specifies the same amino acid; fourfold degenerate codons have the form -N, where N stands for any nucleotide (U, C, A, or G). For example, GUU, GUC, GUA, and GUG all code for valine (Val), which fits the pattern GUN. Note in Table 1.1 that the code for isoleucine is threefold degenerate and those for leucine, arginine, and serine are each sixfold degenerate.

The codons for amino acids are not used randomly in proteins. There are preferred codons for amino acids that differ from one gene to the next and from one organism to another. Codon preferences exist even within redundancy classes. In Drosophila, for example, among codons for histidine, CAC is used more than CAU in a ratio of about 2:1. Similarly, among codons for glutamine, CAG is used more than CAA in a ratio of about 3:1. Another example of nonrandom codon usage is the AUA codon for isoleucine, which tends to be avoided in most proteins in most organisms. In Drosophila, AUA and AUC are used more than AUA in a ratio of about 10:1. One evolutionary

<table>
<thead>
<tr>
<th>TABLE 1.1 THE STANDARD GENETIC CODE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Second nucleotide in codon</strong></td>
</tr>
<tr>
<td>U</td>
</tr>
<tr>
<td>----</td>
</tr>
<tr>
<td>UUU</td>
</tr>
<tr>
<td>UUC</td>
</tr>
<tr>
<td>UUA</td>
</tr>
<tr>
<td>UUG</td>
</tr>
<tr>
<td>C</td>
</tr>
<tr>
<td>CCC</td>
</tr>
<tr>
<td>CUA</td>
</tr>
<tr>
<td>CUG</td>
</tr>
<tr>
<td>A</td>
</tr>
<tr>
<td>AUC</td>
</tr>
<tr>
<td>AUA</td>
</tr>
<tr>
<td>AUG</td>
</tr>
<tr>
<td>G</td>
</tr>
<tr>
<td>GCU</td>
</tr>
<tr>
<td>GUA</td>
</tr>
<tr>
<td>GUG</td>
</tr>
</tbody>
</table>

Note: Codons are nonoverlapping three-base sequences in mRNA, each of which specifies an amino acid in a polypeptide chain or terminates synthesis ("Stop"). The full names of the amino acids are phenylalanine (Phe), leucine (Leu), isoleucine (Ile), methionine (Met), valine (Val), serine (Ser), proline (Pro), threonine (Thr), alanine (Ala), tyrosine (Tyr), histidine (His), glutamine (Gln), asparagine (Asn), lysine (Lys), aspartic acid (Asp), glutamic acid (Glu), cysteine (Cys), tryptophan (Trp), arginine (Arg), and glycine (Gly).
hypothesis that explains the avoidance of AUA is that, because of the degeneracy of the genetic code, the AUA codon might sometimes be translated as AUG, which codes for methionine. Because methionine is likely to change protein structure radically, the mistranslation would be a costly mistake. Through evolutionary time, one by one, the AUA codons in a messenger RNA become replaced with AUU or AUC, minimizing this type of misincorporation error. This misincorporation hypothesis for AUA codon avoidance has not been tested, but it is testable.

**Alleles**

Alternative alleles of a gene differ in their sequence of nucleotides (Figure 1.3). For example, where one allele has a T-A base pair in the DNA, another may have a C-G base pair at the same position. Because of redundancy in the code, not all nucleotide substitutions result in a replacement of one amino acid for another. In Figure 1.3B, for example, if a mutation at the third position in the second codon (asterisk) changes one pyrimidine into the other, the new codon still codes for histidine. On the other hand, some nucleotide substitutions at the third position do result in amino acid replacements. For example, in Figure 1.3C, if the third position in the second codon changes from a pyrimidine to a purine, the codon changes from one for histidine to one for glutamine. Most nucleotide substitutions at codon positions one and two result in amino acid replacements (Figure 1.2D).

Not all alleles differ by a mere nucleotide substitution. Relative to the typical or wildtype allele, some alleles may have a deletion of a number of nucleotide pairs or an insertion into the DNA molecule. The number of nucleotides deleted or inserted may be small (as few as one nucleotide pair) or large. Some insertions are thousands of nucleotide pairs in size. Many large insertions result from the activity of transposable elements, which are specialized sequences of DNA able to replicate and insert at novel positions virtually anywhere in the DNA of the organism in which they are present. Alleles also may differ in the number of copies of short sequences present in tandem arrays in the DNA. For example, many genes in human beings are tandem copies of dinucleotides, such as 5'-CACACACA--; such a repeating sequence is symbolized as (5'-CA-3'). The number of copies (n) of the dinucleotide repeat often range from fewer than ten to hundreds, and the number of copies may differ dramatically from one allele to the next. Some alleles even differ from wildtype in having an inversion of the nucleotide sequence in a region of DNA.

**Genotype and Phenotype**

Within a living cell, genes are arranged in linear order along microscopic threadlike bodies called chromosomes. A typical chromosome may contain several thousand genes. The position of a gene along a chromosome is called the locus of the gene. In most higher organisms, each cell contains two copies of each type of chromosome. Such organisms, in which the chromosomes are present in pairs, are said to be diploid. In each pair of chromosomes, one
member is inherited from the mother through the egg and the other is inherited from the father through the sperm. At every locus, therefore, diploid organisms contain two alleles, one each at corresponding positions in the maternal and paternal chromosomes. If the two alleles at a locus are chemically identical (in the sense of having the same nucleotide sequence along the DNA), the organism is said to be homozygous at the locus under consideration; if the two alleles at a locus are chemically different, the organism is said to be heterozygous at the locus. The term gene is a general term usually used in the sense of locus.

Geneticists make a fundamental distinction between the genetic constitution of an organism and the physical or biochemical attributes of the organism. The genetic constitution of an organism is called the genotype; genotype thus refers to the particular alleles present in an organism at all loci that affect the trait in question. For example, if a trait is influenced by two genes, each with two alleles, then there are nine possible genotypes, as follows.

\[
AA; BB \quad AA; Bb \quad AA; bb
\]
\[
Aa; BB \quad Aa; Bb \quad Aa; bb
\]
\[
aa; BB \quad aa; Bb \quad aa; bb
\]

where \( A \) and \( a \) refer to the alleles of the first gene and \( B \) and \( b \) refer to the alleles of the second gene. In some cases when the alleles are linked (located in the same chromosome), it is sometimes necessary to distinguish between the genotypes \( AB/ab \) and \( Ab/ab \), in which case there are ten possible genotypes.

In contrast to genotype, the physical expression of a genotype is called the phenotype. Examples of phenotypes include hair color, eye color, height, weight, number of kernels on an ear of corn, number of eggs laid by a hen, and round versus wrinkled pea seeds. The distinction between the genetic constitution of an organism (genotype) and the physical or biochemical attributes of the organism (phenotype) is particularly important in cases in which the environment can affect the trait; in such cases, two organisms with the same genotype can nevertheless have different phenotypes because of differences in the environment. Conversely, two organisms with the same phenotype can have different genotypes.

**Problem 1.1** If a gene in a diploid organism has two alternative alleles, show that the number of possible genotypes equals \( m(m + 1)/2 \).

---

**ANSWER:** Consider first the heterozygotes. There are \( m \) ways of choosing the first allele and, having done that, there are \( m - 1 \) ways of choosing a different second allele. Altogether, there are \( m(m - 1)/2 \) different heterozygotes. The division by 2 is necessary because, for each heterozygote—say, \( A_A \)—it makes no difference whether \( A_a \) was chosen first and \( A_a \) second or the other way around. In addition to the heterozygotes, there are \( m \) possible homozygotes. Hence, the total number of diploid genotypes equals \( m(m - 1)/2 + m = m(m + 1)/2 \).

---

**Dominance and Gene Interaction**

Whether each genotype has a single, unique expression of the trait depends on the manner in which the alleles of a gene interact in development. For the alleles of one gene, dominance refers to the concealment of the presence of one allele by the strong phenotypic effects of another. For example, with two alleles there are three possible genotypes:

\[
AA \quad Aa \quad aa
\]

Several types of dominance are distinguished and exemplified in the following examples:

- **Complete dominance:** \( A \) is completely dominant to \( a \) if the phenotypes of \( AA \) and \( Aa \) cannot be distinguished.
- **Incomplete dominance:** \( A \) shows incomplete dominance with respect to \( a \) if the phenotype of \( Aa \) is intermediate between that of \( AA \) and that of \( aa \). This situation is also referred to as partial dominance or intermediate dominance. When the phenotype can be measured on a quantitative scale, for example, the number of kernels on an ear of corn, and the phenotype of \( Aa \) is exactly the average between that of \( AA \) and that of \( aa \), then the alleles are said to be additive alleles and the type of dominance is sometimes called semidominance.
- **Codominance:** \( A \) and \( a \) are codominant if the products of both alleles can be detected in \( Aa \) heterozygotes. Many alleles are codominant at the level of their protein products because two different forms of the polypeptide, encoded by \( A \) and \( a \), can be detected in heterozygotes. At the level of the DNA sequences, all alleles differing in DNA sequence are codominant.
It is important to note that dominance is not a characteristic of alleles so much as a characteristic of the manner in which the phenotype is examined. An allele may show complete dominance if the phenotype is examined in one way, no dominance if examined in another, and codominance if examined in still another. For example, the allele for round pea seeds W studied by Gregor Mendel is completely dominant to that for wrinkled seeds w when the phenotype “round” versus “wrinkled” is examined. The genetic defect in wrinkled seeds is the absence of an enzyme needed for the synthesis of a branched-chain form of starch. Microscopic examination reveals subtle differences in the form of the starch grains in seeds of the three genotypes: WW seeds contain large, well-rounded starch grains, retain water and shrink uniformly as they ripen, so the seeds do not become wrinkled; w seeds lack the branched-chain starch and are irregular in shape because the ripening seeds lose water more rapidly and shrink unevenly. However, heterozygous Ww seeds have starch grains that are intermediate in shape and the seeds shrink uniformly and show no wrinkling. Therefore, at the level of the starch grains, there is incomplete dominance of W and w because the starch grains in the heterozygotes are intermediate between the two homozygotes. Furthermore, the difference in DNA sequence between W and w can readily be detected with modern methods, so that W and w are codominant at the level of DNA sequence.

For traits affected by more than one gene, the relation between genotype and phenotype depends not only on the degree of dominance of the alleles of each gene but also on the type of interaction between the genes in development. For example, suppose that the trait in question is degree of pigmentation and that pigmentation is determined by two alleles of each of two genes, say, A and a and B and b. Suppose further that the total amount of pigmentation in an organism results from the total numbers of A and B alleles present, each of which adds a single unit of pigmentation to the phenotype. Then, as shown in Table 1.2, there are only five possible levels of pigmentation (0 to 4) and genotypes aa BB, Aa Bb, and AA bb all have the same phenotype. Because each uppercase allele adds the same quantity to the total phenotype, the type of gene interaction in Table 1.2 is said to be additive.

### Segregation and Recombination

The essential mechanism of inheritance was established by Gregor Mendel (1822–1884) in experiments with garden peas carried out in the years 1856 to 1863 in a small garden plot next to the monastery in which he lived. Mendel showed that the alleles of each gene segregate from one another in the formation of reproductive cells or gametes. Because of segregation, heterozygous genotypes form equal numbers of gametes containing each allele.

### Table 1.2 A Model of the Additive Gene Action

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Amount of pigmentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>WW</td>
<td>4</td>
</tr>
<tr>
<td>Ww</td>
<td>3</td>
</tr>
<tr>
<td>ww</td>
<td>2</td>
</tr>
<tr>
<td>AA BB</td>
<td>1</td>
</tr>
<tr>
<td>AA bb</td>
<td>0</td>
</tr>
</tbody>
</table>

*At left are shown the nine possible genotypes of two genes with two alleles of each gene. At right is shown the amount of pigmentation expected in each genotype when it is assumed that each allele designated by an uppercase letter is responsible for producing a certain amount of pigment.

*Measured as an increase in pigmentation over that in an AA BB genotype.

Furthermore, because gametes unite at random in fertilization, the following are the results of simple Mendelian segregation:

- **AA x AA** matings produce all **AA** progeny.
- **AA x Aa** matings produce 1/2 **AA** and 1/2 **Aa** progeny.
- **AA x aa** matings produce all **Aa** progeny.
- **Aa x Aa** matings produce 1/4 **AA**, 1/2 **Aa**, and 1/4 **aa** progeny.
- **Aa x aa** matings produce 1/2 **Aa** and 1/2 **aa** progeny.
- **aa x aa** matings produce all **aa** progeny.

The physical basis of Mendelian segregation is that the maternal and paternal pairs of chromosomes are segregated into different cells in the formation of gametes. Prior to their separation, the maternal and paternal chromosomes associate intimately all along their length and alleles may be interchanged in the process of recombination (Figure 1.4). The interchange of parts takes place after the chromosomes have replicated, and only two of the four chromosome strands participate in any one exchange. Recombination results in the creation of allele combinations different from either parental chromosome. In Figure 1.4, the A and b and a and B combinations are recombinant, whereas the A and a and B and b combinations are parental (nonrecombinant). Therefore, a single exchange between parental chromosomes results in two recombinant and two nonrecombinant gametes.

In organisms with an XX–XY chromosomal mechanism of sex determination, Mendelian segregation randomizes the sex ratio at fertilization. In mammals and many other animals, sex is determined by sex chromosomes: males have an X and a Y chromosome, and females have two X chromosomes. In males, the X and Y chromosomes segregate, yielding equal proportions of
X-bearing and Y-bearing sperm. If both types of sperm are equally able to fertilize eggs, then random union of sperm with eggs yields \( \frac{1}{2}XX \) (female) and \( \frac{1}{2}XY \) (male) chromosome constitutions.

**PROBABILITY IN POPULATION GENETICS**

The basic concepts of probability needed for elementary population genetics are quite straightforward. They will be introduced with the concrete example of genetic segregation in Figure 1.5, which deals with the progeny of the mating

\[
(A) \quad \text{Addition rule}
\]

**Mating**

\[ Aa \times Aa \]

**Offspring**

\[ \frac{1}{4}AA + \frac{1}{2}Aa + \frac{1}{4}aa \]

\( A- \) means "Offspring either AA or Aa"

\[ P(A-) = \frac{1}{4}(AA) + \frac{1}{2}(Aa) = \frac{3}{4} \]

**(B) Multiplication rule**

<table>
<thead>
<tr>
<th>Sibship</th>
<th>Birth Order</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A- A- A-</td>
<td>( \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8} )</td>
</tr>
<tr>
<td>2</td>
<td>A- A- aa</td>
<td>( \frac{1}{2} \times \frac{1}{2} \times \frac{1}{4} = \frac{1}{8} )</td>
</tr>
<tr>
<td>3</td>
<td>A- aa A-</td>
<td>( \frac{1}{2} \times \frac{1}{4} \times \frac{1}{2} = \frac{1}{8} )</td>
</tr>
<tr>
<td>4</td>
<td>aa A- A-</td>
<td>( \frac{1}{4} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8} )</td>
</tr>
<tr>
<td>5</td>
<td>A- aa aa</td>
<td>( \frac{1}{2} \times \frac{1}{4} \times \frac{1}{4} = \frac{1}{8} )</td>
</tr>
<tr>
<td>6</td>
<td>aa A- aa</td>
<td>( \frac{1}{4} \times \frac{1}{2} \times \frac{1}{4} = \frac{1}{8} )</td>
</tr>
<tr>
<td>7</td>
<td>aa aa A-</td>
<td>( \frac{1}{4} \times \frac{1}{4} \times \frac{1}{2} = \frac{1}{8} )</td>
</tr>
<tr>
<td>8</td>
<td>aa aa aa</td>
<td>( \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} = \frac{1}{8} )</td>
</tr>
</tbody>
</table>

**Figure 1.4** Recombination results from a physical interchange of parts between chromosomes. New combinations of alleles are created that differ from either parental chromosome. The physical interchange of parts takes place in gamete formation after the chromosomes have replicated, and only two of the four chromosome strands participate in any one exchange.

The basic concepts of probability illustrated by Mendel's segregation in the mating \( AA \times Aa \). The elementary outcomes of the mating are the possible genotypes of each progeny—\( AA, Aa, \) and \( aa \)—and these are realized with probabilities \( \frac{1}{4}, \frac{1}{2}, \) and \( \frac{1}{4} \), respectively. (A) The compound event \( A- \) consists of the two elementary outcomes \( AA \) and \( Aa \), and the probability of \( A- \) is the sum of the probabilities of these elementary outcomes (addition rule). (B) The possible distributions of genotypes \( A- \) and \( a \) in sibs of six three offspring. Successive births are independent, and so the probability of any sibship equals the product of the probabilities for each birth separately (multiplication rule).
As x Aa. Considerations in probability always begin with an experiment of some kind. The experiment may be either a real experiment or a conceptual experiment. In Figure 1.5, it is a conceptual experiment in which Aa is crossed with Aa. In probability calculations, it is also necessary to define all possible outcomes of the experiment. The outcomes are called elementary outcomes because they are defined in such a way that, in any repetition of the experiment, one and only one of the elementary outcomes must be realized. For example, if we are interested in the genotypes among the progeny of the mating Aa x Aa, the possible elementary outcomes for each offspring are either AA, Aa, or aa. (Note that, in defining these as the elementary outcomes, we are ignoring the possibility of either A or a mutating to a novel allele.) To proceed further, we must assign to each elementary outcome a probability, a number between 0 and 1 that measures how much confidence we have that the outcome will be realized. The probabilities assigned to the outcomes are based on genetic reasoning, intuition, or experience. One requirement of the assigned probabilities is that the probabilities of all the elementary outcomes must add to 1, this is the mathematical consequence of requiring that one of the elementary outcomes must be realized. For example, if there are three elementary outcomes, and all are equally probable, then each has a probability of 1/3. In Figure 1.5, the probabilities assigned to the elementary outcomes AA, Aa, and aa are 1/4, 1/2, and 1/4, respectively, because these are the relative proportions of the three progeny genotypes expected from Mendelian segregation.

The Addition Rule

An outcome of a conceptual experiment is an event. The distinction between an event and an elementary outcome is that an event can include more than one elementary outcome. For example, in Figure 1.5A, the event “the offspring has at least one copy of the dominant A allele” consists of two elementary outcomes, namely, genotypes AA and Aa. This event may be symbolized A−, where the dash indicates that the unspecified allele may be either A or a. For events defined in terms of elementary outcomes, the probability of an event equals the sum of the probabilities of the elementary outcomes included in the event. In the present example,

\[ \Pr(A-) = \Pr(AA) + \Pr(Aa) = \frac{1}{4} + \frac{1}{2} = \frac{3}{4} \]

More generally, two events are mutually exclusive if they cannot be realized simultaneously. The addition rule states that, for mutually exclusive events, the probability that either one or the other is realized equals the sum of the probabilities of the separate events.

The Multiplication Rule

Figure 1.5B shows all possible genotypes of sibships of three offspring from the mating Aa x Aa, with each offspring classified as A− versus a. (A sibship is a group of brothers and sisters.) The probability of A− in any particular birth is 3/4 and that of a is 1/4. The probabilities at the right are the overall probabilities for each of the sibships. They are obtained by multiplication of the probability for each birth because successive births are independent, which means that the genotype of any birth has no effect on the genotype of any other birth. Because of the independence, among the 3/4 of the sibships with A− in the first birth, 3/4 will have A− in the second birth, and among the 3/4 x 3/4 of the sibships with A− in the first two births, 3/4 will have A− in the third birth. Therefore, the overall probability of three A− births is 3/4 x 3/4 x 3/4. The reasoning for the other types of sibships is similar. More generally, the multiplication rule states that, whenever two events are independent, the probability of their joint realization is the product of the probabilities of their being realized separately.

Repeated Trials

The sibships in Figure 1.5B are an example of repeated trials of a conceptual experiment. Repeated trials are encountered frequently in probability. They govern tosses of a coin or dice, deals of cards, successive spins of a roulette wheel, and so forth. Repeated trials are also important in population genetics because successive offspring of a mating are independent events and thus repeated trials. Furthermore, it is apparent from Figure 1.5B that the different birth orders are mutually exclusive; any sibship can have one and only one birth order of A− or a. Because the birth orders are mutually exclusive, their probabilities may be combined by the addition rule. Hence, the composite events below have the following probabilities:

\[ \Pr(\text{two } A- \text{ and one } \text{a}) = 3/4 \times 3/4 \times 1/4 = 9/64 \]

\[ \Pr(\text{one } A- \text{ and two } \text{a}) = 3/4 \times 3/4 \times 1/4 = 9/64 \]

Note in Figure 1.5B that, when the sibships with the same number of A− and a genotypes are combined, the overall probabilities are given by successive terms in the expansion of:

\[ \left( \frac{3}{4} A- + \frac{1}{4} \text{a} \right)^3 = 1 \times \left( \frac{3}{4} \right)^3 A- A- A- \\
+ 3 \times \left( \frac{3}{4} \right)^2 \left( \frac{1}{4} \right) A- A- \text{a} \\
+ 3 \times \left( \frac{3}{4} \right) \left( \frac{1}{4} \right)^2 \text{a} \text{ a} A- \\
+ 1 \times \left( \frac{1}{4} \right)^3 \text{a} \text{ a} \text{ a} \]

The coefficients 1 : 3 : 3 : 1 are the number of combinations in which each triad of genotypes can be born: 1 for A− A− A−, 3 for A− A− a, and 3 for A− a a (because the aa genotype can be born either first, second, or third), and so forth. Each power of 3/4 and 1/4 is the probability that any one of the birth orders will be realized,
for example, \( P(A')P(B') \) is the probability that any sibship with two A- and one a/at genotype will be realized.

In all cases of repeated and independent trials, the overall probabilities are given by analogous expansions. Suppose that any one trial may result in either of two mutually exclusive events, A or B, and that the probability of event A is \( p \) and that of event B is \( q \) (with \( p + q = 1 \)). Among a total of \( n \) independent trials, what is the probability that A is realized exactly \( r \) times and B is realized exactly \( n - r \) times? By the multiplication rule, any particular combination of \( r \) As and \( n - r \) Bs has a probability \( p^rq^{n-r} \). Deducing the total number of combinations of \( r \) As and \( n - r \) Bs is a little less obvious, but it is given by the coefficient of the term \( p^rq^{n-r} \) in the expansion of \( (p + q)^n \), which equals

\[
\frac{n!}{r!(n-r)!} \quad 1.1
\]

where the exclamation point means the factorial, the product of all integers from 1 through the number in question. For example, \( n! = 1 \times 2 \times 3 \times \cdots \times n \). For consistency, the number 0! is defined as \( 0! = 1 \).

Equation 1.1 is often called a binomial coefficient because it arises in the expansion of the two forms \((p + q)^n\). To understand the reason why Equation 1.1 yields the correct number of combinations of \( r \) As and \( n - r \) Bs, first consider what the \( n! \) means. It is the total number of ways that any set of \( n \) objects can be arranged in order. There are \( n \) ways to choose the first object and, having chosen the first, \( n - 1 \) ways to choose the second and, having chosen the first two, \( n - 2 \) ways to choose the third, and so on, yielding \( n \times (n - 1) \times (n - 2) \times \cdots \times 1 = n! \). Furthermore, for each arrangement of \( n \) objects of which \( r \) are As and \( n - r \) are Bs, there are \( r! \) ways to arrange the As among themselves and \( n - r \) ways to arrange the Bs among themselves, for a total of \( r! \times (n - r)! \) arrangements. Because each of the \( n! \) combinations of \( r \) As and \( n - r \) Bs includes \( r! \times (n - r)! \) equivalent arrangements of the As and Bs, the total number of different arrangements of \( r \) As and \( n - r \) Bs equals the ratio given in Equation 1.1.

Equation 1.1 gives the number of different arrangements of \( r \) As and \( n - r \) Bs. Each arrangement has a probability given by \( p^rq^{n-r} \). Therefore, using the addition rule, the probability that \( r \) repeated trials yields \( r \) realizations of A and \( n - r \) realizations of B equals

\[
\frac{n!}{r!(n-r)!} \quad 1.2
\]

As an example of the use of Equation 1.2, consider the probability that a sibship of 12 offspring from the mating \( Aa \times Aa \) perfectly matches the expected Mendelian ratio of 9 A- and 3 a/a. In this case, \( p = \frac{1}{2} \), \( q = \frac{1}{2} \), \( n = 12 \), \( r = 9 \), and \( n - r = 3 \). The required probability from Equation 1.2 is therefore

\[
\frac{12!}{9!3!} \left( \frac{1}{2} \right)^9 \left( \frac{1}{2} \right)^3 = 220 \times 0.00751 \times 0.000005 = 0.258
\]

The implication of this calculation is that, whereas the "expected" ratio is 9 A- : 3 a/a, only a little more than 25% of such sibships actually have the expected distribution.

**PROBLEM 1.2.** Suppose that a society decided to limit the number of males by passing a law denying further reproduction to any woman who gives birth to a male child. Given a ratio of males to females at birth of 1 : 1, how would such a law affect the sex ratio? Suppose further that, in practice, any woman who has a female child voluntarily terminates further reproduction with probability \( p \). In this case, what is the proportion of males in sibships of size \( n \)?

**ANSWER.** The law would have no effect on the sex ratio. To understand why, consider the first birth across the entire population. The sex ratio among these offspring must be 50% males. Consider now the second birth. The sex ratio among these offspring must also be 50% males. Indeed, the sex ratio in any birth must be 50% males, and so too is the sex ratio in the population of births as a whole. In regard to the second part of the problem, note that sibships of size \( n \) can be separated into two classes: those in which the final birth is a male (and the mother's further reproduction is denied) and those in which the final birth is a girl (in which the mother voluntarily stops reproducing with probability \( p \)). These types of sibships occur in the ratio \( \frac{1}{2} : \frac{1}{2} \), which means in the proportions \( 1/(1 + q) \) and \( p/(1 + p) \), respectively. The first type of sibship has a proportion of males of 1/n and the second has a proportion of males of 0. Hence, the proportion of males as a function of sibship size equals \( (1/n) \times [1/(1 + p)] + 0 \times [p/(1 + p)] = 1/n(1 + p) \). Note that, for \( p = 0 \), the proportion of males as a function of sibship size decreases according to the series \( 1/2, 1/3, 1/4, \ldots \). Nevertheless, the sex ratio in the population as a whole equals \( \frac{1}{2} \) for this and any other value of \( p \).
PHENOTYPIC DIVERSITY AND GENETIC VARIATION

One of the universal attributes of natural populations is that organisms differ in phenotype with respect to many traits. Phenotypic diversity in many traits is impressive even with the most casual observation. Among human beings, for example, there is diversity with respect to height, weight, body conformation, hair color and texture, skin color, eye color, and many other physical and psychological attributes or skills. Population geneticists must deal with this phenotypic diversity, and especially with that portion of the diversity that is caused by differences in genotype. In particular, the field of population genetics has set for itself the tasks of determining how much genetic variation exists in natural populations and of explaining its origin, maintenance, and evolutionary importance. Genetic variation, in the form of multiple alleles of many genes, exists in most natural populations. In most sexually reproducing populations, no two organisms (barring identical twins or other multiple identical births) can be expected to have the same genotype for all genes. Thus, it becomes important to describe how alleles in natural populations are organized into genotypes—to determine, for example, whether alleles of the same or different genes are associated at random.

Allele Frequencies in Populations

Much of the phenotypic variation in natural populations does not yield simple Mendelian segregation ratios such as 1:1 or 3:1 in pedigrees. Some differences in phenotype are environmental in origin and are not expected to show Mendelian segregation. However, simple Mendelian segregation is not usually observed even for traits whose expression is influenced more or less strongly by genetic factors. Although the underlying genetic factors do segregate in pedigrees in Mendelian fashion, the segregation is concealed by several complications. First, environmental effects on the trait may be strong enough to mask the genetic segregation. Second, genetic effects on many traits are determined by the joint effects of the alleles of two or more genes, and the segregation of any one gene in a pedigree may be obscured by the segregation of others.

On the other hand, some phenotypic diversity in populations does show simple Mendelian segregation. In the snapdragon Antirrhinum majus, for example, whether the flower color is red, pink, or white is determined by the alleles l and i of a single gene. The genotypes ll, li, and ii have red, pink, and white flowers, respectively, an example of incomplete dominance.

Populations containing both the l and i alleles will include plants whose flowers are red (ll), pink (li), or white (ii) in proportions determined by the allele frequencies of the l and i alleles in the population as well as by the manner in which the alleles are united in fertilization. By the allele frequency of a specified allele, we mean the proportion of all alleles of the gene that are of the specified type. To take a hypothetical example, suppose 400 members of a population were classified as to flower color and the finding was: 165 red, 190 pink, and 45 white. Because the flower color reveals the genotype, we may infer that the sample of 400 includes 165 ll, 190 li, and 45 ii genotypes. The observed numbers of l and i alleles are therefore:

\[
\begin{align*}
    l & : 2 \times 165 + 190 = 520 \\
    i & : 190 + 2 \times 45 = 280
\end{align*}
\]

The factors of 2 are included for the homozygous genotypes because each ll genotype contains two l alleles and each ii genotype contains two i alleles. The total number of alleles in the sample equals 2 \times 400 = 800. Therefore, if we let \( p \) represent the frequency of the \( l \) allele and \( q \) represent the frequency of the \( i \) allele (with \( p + q = 1 \) because these are the only alleles of the gene in question), then we can estimate \( p \) and \( q \) from the observations as:

\[
\begin{align*}
    \hat{p} & = \frac{520}{800} = 0.65 \\
    \hat{q} & = \frac{280}{800} = 0.35
\end{align*}
\]

Note that, if the l and i alleles were combined into genotypes at random, the expected frequencies of three genotypes can be calculated from the rule for repeated trials by expanding the binomial \( p + q = p^2 + 2pq + q^2 \) into:

\[
\begin{align*}
    ll & : (0.65)^2 \times 400 = 169 \\
    li & : 2 \times 0.65 \times 0.35 \times 400 = 182 \\
    ii & : (0.35)^2 \times 400 = 49
\end{align*}
\]

Hence, the observed numbers in this hypothetical population are very close to those expected with random combinations of alleles. The proportions \( p^2, 2pq, \) and \( q^2 \) for the three genotypes when two alleles are combined at random constitute the Hardy-Weinberg principle, which is one of the basic principles in population genetics. The Hardy-Weinberg principle is discussed in detail in Chapter 2.

**Problem 1.3** Suppose that a random sample of 400 snapdragons from a population includes 185 red, 150 pink, and 65 white. Estimate the allele frequency of \( p \) of \( l \) and \( q \) of \( i \). Assuming random combinations of alleles in the genotypes, what are the expected numbers of
the three genotypes? Do the observed data seem to fit the expectations?

ANSWER  Among the total of 800 alleles, the observed number of \( I \) alleles is \( 2 \times 185 + 150 = 520 \) and that of \( i \) alleles is \( 150 + 2 \times 65 = 280 \). Therefore, \( \hat{p} = 520/800 = 0.65 \) and \( \hat{q} = 280/800 = 0.35 \). Note that the estimated allele frequencies are the same as above, even though the observed numbers of the genotypes are different. With random combinations of alleles in the genotypes, the expected numbers are again 169 red, 182 pink, and 49 white. Compared to the observations, there appear to be too many homozygous genotypes and too few heterozygous genotypes. (A statistical method for deciding whether the fit is satisfactory or not is discussed in Chapter 2.)

Parameters and Estimates
In the discussion of flower color in snapdragons, we made a subtle distinction between the actual allele frequency of the \( I \) allele (designated \( p \)) and the estimated allele frequency of the \( I \) allele (symbolized \( \hat{p} \)). The distinction is necessary whenever an experimenter makes inferences about an entire population from an examination of a random sample from the population. Quantities used in describing entire populations are parameters. In the snapdragon example, the parameter of interest is the allele frequency \( p \) of \( I \) in the entire population. Because we only have access to a sample of 400 organisms from the population, the true value of \( p \) is unknown. The best we can do is make an estimate of \( p \) based on a sample, hoping that the sample is representative of the population as a whole. The estimate obtained from the sample is designated \( \hat{p} \) to emphasize that it is an estimate rather than the true value. In this book, whenever it is necessary to distinguish parameters from their estimates, we use unencumbered symbols for parameters (for example \( p \) for the unknown frequency of an allele in a specified population) and the same symbol with a circumflex for the estimated value (in this example \( \hat{p} \)).

The Standard Error of an Estimate
The distinction between a parameter and an estimate is important because different samples may yield different values of the estimate for the same reason that different siblings may yield different segregation ratios, namely, chance variation from one repeated trial to the next. The estimation of an allele frequency can be treated as repeated trials by supposing that the alleles are sampled at random, one by one, from a very large population. In the snapdragon example, there are 800 alleles sampled. If the allele frequency of \( I \) has the true value \( p = 0.65 \), then the repeated-trials interpretation implies that all possible outcomes of 800 trials have probabilities given by successive terms in the expansion of \( (0.65 i + 0.35 i)^{800} \). This is not an expansion that one would want to do by hand, but the binomial expression makes evident the underlying random-sampling process that accounts for variation in the estimate of \( \hat{p} \) from one sample of 800 alleles to the next.

Unless \( p \) is quite close to 0 or quite close to 1, there is a convenient approximation to the binomial expansion \( (p + q \hat{p})^n \), where \( n \) is the number of alleles sampled. As \( n \) becomes large, the distribution of \( \hat{p} \) approaches the familiar bell-shaped curve called the normal distribution. The normal distribution features prominently in the analysis of traits determined jointly by multiple genetic and environmental factors and it is discussed in detail in that context (Chapter 9). For present purposes, it is sufficient to note that the degree to which the values of \( \hat{p} \) are clustered around the overall average depends on a quantity called the standard error:

\[
\text{SE} = \sqrt{\frac{\hat{p} \hat{q}}{n}}
\]

where \( \hat{q} = 1 - \hat{p} \). If the sampling and estimation of \( p \) were repeated many times using the same population, then the values of \( \hat{p} \) would be expected to be clustered symmetrically around \( p \) according to the standard error as follows:

- Approximately 68% of the estimates \( \hat{p} \) lie within plus or minus one standard error of \( p \).
- Approximately 95% of the estimates \( \hat{p} \) lie within two standard errors of \( p \).
- Approximately 99.7% of the estimates \( \hat{p} \) lie within three standard errors of \( p \).

To put the matter in another way, with repeated sampling, 32% of the estimates would be expected to differ from the true value by more than one standard error, 5% by more than two standard errors, and only 0.3% by more than three standard errors.

As an illustration of the variation among repeated estimates of \( p \), Figure 1.6 shows the values of \( \hat{p} \) obtained in 100 repetitions of the experiment of sampling 800 alleles from a large population in which the true allele frequency is \( p = 0.65 \). Each of the 100 samples was created by computer simulation using a random-number generator that yielded a 1 with probability 0.65 and a 0 with probability 0.35. For each sample of 800, therefore, the estimate \( \hat{p} \) equals the number of 1s in the sample divided by 800. As is evident in Figure 1.6, the distribution of \( \hat{p} \) values is more or less bell-shaped but not
Figure 1.6  Estimates of allele frequency based on 100 samples, each of size 400 diploid organisms, from a population in which the actual allele frequency is 0.65. The standard error equals 0.017, and the distribution of the estimates is very close to the bell-shaped distribution expected theoretically.

The scale across the top gives the ranges of the estimates as multiples of the standard error.

Exactly so because it is based on only 100 samples rather than an infinite number. The overall mean \( \hat{p} \) from all 100 samples combined (80,000 observations) equals 0.6492, which is very close to the true value of \( p \). Furthermore, the distribution of the estimates fits the predictions based on the standard error quite well.

To apply Equation 1.3 to the data in Figure 1.6, note first that \( \hat{p} = 0.65 \) with \( n = 800 \), and so \( s \) in Equation 1.3 equals \( \sqrt{(0.65 \times 0.35)/800} = 0.017 \). Because 68% of the samples are expected to yield values of \( \hat{p} \) in the range \( p \pm s \), and because the expected distribution is symmetrical, 34 of the values in Figure 1.6 are expected in the range \( p - s \) to \( p + s \) (0.633–0.650) and 34 in the range \( p + 2s \) to \( p + 3s \) (0.650–0.667); the actual numbers are 33 in the first interval and 35 in the second. By the same reasoning, 95% of the values should lie in the range \( p \pm 2s \), or 47.5% on each side of the mean; because 34% of the values on either side of the mean are in the range \( p \pm s \), the implication is that 47.5–34 = 13.5% of the values should lie in the range \( p \pm 2s \) to \( p \pm 3s \) and 13.5% should lie in the range \( p \pm s \) to \( p + 2s \). For the data in Figure 1.6, these ranges are 0.616–0.633 and 0.667–0.684; the actual number in each interval is 18 and 10, respectively, as against the theoretical 13.5 in each. Likewise, the standard error predicts that 0.3% of the samples will deviate by more than 3\( s \) from the mean, as compared with the observed 2.

Estimates and their standard errors are often presented as \( \hat{p} \pm s \), or 0.65 ± 0.017 in the present example. The 68%, 95%, and 99.7% cutoffs for \( \pm 1, \pm 2, \) and \( \pm 3 \) standard errors provide one manner in which the reliability of an estimate may be interpreted. Estimates may also be presented alternatively in terms of a range called a confidence interval, which expresses a degree of confidence that the true value of a parameter lies in some specified interval. The most frequently encountered confidence interval is the 95% confidence interval, defined as the interval \( (\hat{p} - 2s, \hat{p} + 2s) \). Because 95% of repeated samples are expected to yield estimates in a range \( \pm 2s \) around the true mean, then 95% of the time the interval \( (\hat{p} - 2s, \hat{p} + 2s) \) is expected to include the true value of the parameter \( p. \)

In the snapdragon example with \( \hat{p} = 0.65 \) and \( s = 0.017 \), the 95% confidence interval is 0.616–0.684.

PROBLEM 1.4  The MN blood groups in human beings are determined by two alleles of a single gene, designated \( M \) and \( N \). Each allele results in the production of a different type of polysaccharide molecule on the surface of red blood cells, which can be distinguished by means of appropriate chemical reagents. The types of molecules corresponding to the \( M \) and \( N \) alleles are designated \( M \) and \( N \), respectively. The \( M \) and \( N \) alleles are codominant; that is, genotype \( MM \) produces only the \( M \) substance and has blood group \( M \), genotype \( NN \) produces only the \( N \) substance and has blood group \( N \), and the heterozygous genotype \( MN \) produces both the \( M \) and \( N \) substances and has blood group \( MN \). Among a sample of 1000 British people (Race and Sanger 1975), the observed numbers of each blood group were 298 \( M \), 489 \( MN \), and 213 \( N \). Using these data, estimate the allele frequency \( p \) of the \( M \) allele and calculate its standard error. What are the 68%, 95%, and 99.7% confidence intervals for \( p \)?

ANSWER  Because each genotype has a unique phenotype, the sample contains \( 2 \times 298 + 489 = 1085 \) \( M \) alleles, and so \( \hat{p} = 1085/2000 = 0.5425 \). The standard error \( s = \sqrt{(0.5425)(0.4575)/2000} = 0.0111 \). The 68%, 95%, and 99.7% confidence intervals for \( p \) are \( \hat{p} \pm 1s, \hat{p} \pm 2s, \) and \( \hat{p} \pm 3s \), respectively, and so the confidence intervals are 0.5202–0.5647 (95%), and 0.5092–0.5758 (97.5%).
MODELS IN POPULATION GENETICS

Population geneticists must contend with factors such as population size, patterns of mating, geographical distribution of organisms, mutation, migration, and natural selection. Although we wish ultimately to understand the combined effects of all these factors and more, the factors are so numerous and interact in such complex ways that they cannot usually be grasped all at once. Simpler situations are therefore devised, situations in which a few identifiable factors are the most important ones and others can be neglected. An intentional simplification of a complex situation is a model. There are several types of models, each designed to eliminate extraneous detail in order to focus attention on the essentials. Some models are experimental. An experimental model may consist of a laboratory experiment with population cages of *Drosophila* or growing cultures of bacteria. An experimental model may also consist of observations of natural populations in particular locations or at particular times in which evolutionary forces of interest may be presumed to be present. Models of this type include the study of the origin and spread of insecticide resistance in insects or antibiotic resistance in bacteria.

A model may also be a conceptual simplification. Conceptual models have a number of uses. They require a concise statement of presumed mechanisms and interactions; they afford a framework for interpreting observations and setting research priorities; they enable extrapolation into the future or beyond the range of known parameters; and they suggest tests of consistency between theory and observation.

A conceptual model may consist of verbal arguments logically linking a chain of hypothesis and deductions. Another type of conceptual model is a computer program that simulates the random component in a process or that calculates the values of changing quantities in a complex system based on prescribed numerical relations. An example of a computer model is the one for examining the result of repeated random sampling whose outcome is depicted in Figure 1.6. In population genetics, a kind of model frequently encountered is a mathematical model, which is a set of hypotheses that specifies the mathematical relations between measured or measurable quantities (the parameters) in a system or process. Mathematical models can be extremely useful:

- They express concisely the hypothesized quantitative relationships between parameters.
- They reveal which parameters are the most important in a system and thereby suggest critical experiments or observations.
- They serve as guides to the collection, organization, and interpretation of observed data.
- They make quantitative predictions about the behavior of a system that can, within limits, be confirmed or shown to be false.

The validity of any model must be tested by determining whether the hypotheses on which it is based and the predictions that grow out of it are consistent with observations.

A mathematical model is always simpler than the actual situation it is designed to elucidate. A model is supposed to be simple if it is not simpler than the real situation, then it isn't a model. Models are simpler than real situations because many features of real life are intentionally ignored. To include every aspect of a complex system would make a model too complex and unwieldy. Construction of a model always requires a compromise between realism and manageability. A completely realistic model is likely to be too complex to handle mathematically, and a model that is mathematically simple may be so unrealistic as to be useless. Ideally, a model should include all essential features of the system and exclude all nonessential ones. How good or useful a model is often depends on how closely this ideal is approximated. In short, a model is a sort of metaphor or analogy. Like all analogies, it is valid only within certain limits but, when pushed beyond these limits, becomes misleading or even absurd.

In this book, we are going to take many liberties with mathematical rigor. Our excuse is that the basic ideas of a model are often obscured rather than illuminated by excessive attention to mathematical detail. Our authority for the approach is the great physicist Richard Feynman, who wrote in one of his books:

Mathematicians may be completely repelled by the liberties taken here. The liberties are taken not because the mathematical problems are considered unimportant. On the contrary, I hope to encourage the study of these forms from a mathematical standpoint. In the meantime, just as a poet has a license from the rules of grammar and pronunciation, we should like to ask for "physical license" from the rules of mathematics in order to express what we wish to say in as simple a manner as possible.

**Exponential Population Growth**

To illustrate the nature of mathematical models (as well as some of their limitations) we consider the dynamics of population growth, a subject of considerable interest in population genetics and population biology. In Figure 1.7, the solid dots show the increase in the number of cells of the yeast *Saccharomyces cerevisiae* in a defined quantity of culture medium. The number of cells increases slowly at first (0-4 hours), then more rapidly (hours 4-12), then more slowly again (hours 12-18). As a first approximation of the early stages of population growth, we may assume that a constant fraction
of the cells reproduces in each interval of time. To simplify matters further, we will assume that the population size does not change gradually but changes in a discrete and instantaneous "jump" at the end of each hour. A model of this type is a discrete model of population growth. Thus, we may write

\[ N_t = N_{t-1} + rN_{t-1} \]  

where \( N_t \) and \( N_{t-1} \) represent population size at the end of hours \( t \) and \( t - 1 \) and where \( r \) is a constant called the intrinsic rate of increase equal to the fraction of cells that reproduce in each interval of time. This equation says that the population size at the end of hour \( t \) is the sum of two components: (1) all the cells present at the end of hour \( t - 1 \) (which means that none of the cells die), and (2) the progeny of the \( rN_{t-1} \) cells that divided in the interval.

Equation 1.4 illustrates a feature of theoretical population genetics that sometimes leads to confusion: the same symbols are often used for different things. In this equation, \( r \) is the intrinsic rate of increase in population number. In other equations in population genetics, \( r \) is the recombination fraction between two genes linked in the same chromosome. The symbol \( r \) is used for still other parameters also. Any possible confusion could be avoided by indicating each parameter with a different letter; this solution is impractical because it quickly runs out of letters, even including Greek letters. Another way is to distinguish different meanings of the same letter by typography, the use of superscripts, subscripts, and so forth. The problem with this approach is that even simple equations get to look imposing. Still another solution, which is the one adopted in this book, is to ask the reader to play close attention to the context so that, for example, \( r \) as used in the context of population growth is not confused with \( r \) used in the context of genetic linkage and recombination.

The solution to Equation 1.4 is straightforward. Because \( N_t = (1 + r)N_{t-1} \), it follows that \( N_{t+1} = (1 + r)N_t \). Consequently, we can write \( N_t = (1 + r)(1 + r)N_{t-1} = (1 + r)^2N_{t-2} \). However, \( N_{t+1} = (1 + r)N_t \), and so \( N_t = (1 + r)^{t-1}N_0 \). Continuing in this manner, we eventually deduce that

\[ N_t = (1 + r)^tN_0 \]

For the data in Figure 1.7, if we set \( N_0 = 10 \) (the observed number) and \( r = 0.7083 \), the first few points from Equation 1.5 (indicated by crosses) fit very well — \( N_1 = 10, N_2 = 17, N_3 = 29, N_4 = 50 \). Then the model starts to break down: \( N_5 = 85, N_6 = 145, N_7 = 249 \), and thereafter the fit becomes very bad indeed. The lesson from this example is that many models have a range over which they are reasonable approximations to the real world, in this case, for a short time after a yeast culture is inoculated. If the model is extrapolated beyond its range of validity, it yields nonsense. The problem for many models in population genetics is that their range of validity is unknown.

In Equation 1.5, \( N \) is defined only for \( t \) equal to positive integers because of the discrete nature of the model. Population growth is actually a continuous process. Population size increases gradually rather than in jumps. The continuous-growth version of Equation 1.5, shown by the dashed line labeled "exponential curve" in Figure 1.7, is given by

\[ N(t) = N(0)e^{rt} \]

where \( n_0 = \ln(1 + r) \). The rationale for Equation 1.6 is based on the same sort of argument as Equation 1.4 but compressing the time scale. Whereas Equation 1.4 assumes that each unit of time is one hour, suppose that each time unit were, say, one minute. In slowing down the time scale in this manner, we must also decrease the value of \( r \), otherwise too many organisms would reproduce in each unit of time. Therefore, by analogy with Equation 1.4, we can write \( N_t - N_{t-1} = r(N_{t-1}) \), but here \( r \) is the intrinsic rate of increase in the new time scale. If \( N(t) \) is a smooth, continuous function and not changing too fast, then it is easy to convince yourself that \( N_t - N_{t-1} \) should approximate the derivative of \( N(t) \), which is the change in \( N(t) \) in a small
interval of time, and that $N_{t-1}$ should be close to $N(t)$ because we have assumed that $N(t)$ is not changing very fast in the new time scale. Therefore, we can write

$$\frac{dN(t)}{dt} = \eta_0 N(t)$$

or

$$\frac{dN(t)}{N(t)} = \eta_0 dt$$

Because $\ln N(t) = \int \frac{dN(t)}{N(t)}$, where $\ln$ is the base of natural logarithms, the solution of Equation 1.8 is $N(t) = r_0 + C$, where $C$ is a constant chosen so that $N(t) = N(0)$ when $t = 0$. (Hence, $C = \ln N(0)$.) Expressing the solution in terms of $N(t)$ rather than $\ln N(t)$ yields Equation 1.6. Furthermore, comparing Equation 1.6 with Equation 1.5, it is clear that

$$N(t) = (1 + \frac{r}{r_0}) N_0$$

and therefore $r_0 = \ln(1 + r)$ is the relation between the parameter $r_0$ in the continuous model and the parameter $r$ in the discrete model. Equation 1.6 is the exponential function plotted in Figure 1.7 with $N(0) = 10$ and $r_0 = 0.53835$.

PROBLEM 1.5 Under optimal culture conditions, the bacterium *Escherichia coli* can double in population size every 20 minutes. Because population growth is continuous, Equation 1.7 is the appropriate model. A single cell of *E. coli* is cylindrical in shape and has a volume of approximately $1.6 \times 10^{-12}$ cm$^3$. A standard soccer ball has a diameter of 22 cm (roughly 9 inches) and a volume of approximately 5600 cm$^3$.

(a) What intrinsic rate of increase $r_0$ per minute results in a doubling time of 20 minutes?
(b) Starting with a single cell of *E. coli* growing under optimal conditions, how long would it take to produce enough cells to fill one soccer ball?
(c) How many soccer balls could be filled with cells after 24 hours of unrestricted growth?

**ANSWER**

(a) Set $N(20) = 2N(0) = N(0) \exp (r_0 \times 20)$, where $\exp (\cdot)$ stands for $e^\cdot$. Therefore, $r_0 = (\ln 2)/20 = 0.034657$. (b) One soccer ball full of cells equals $5600/(1.6 \times 10^{-12}) = 3.5 \times 10^{25}$ cells. The time needed to produce this many cells is given by $t = \ln (3.5 \times 10^{25})/r_0 = 1032.7$ minutes (17.2 hours). (c) After 24 hours (1440 minutes) of unrestricted growth, one cell yields $\exp (r_0 \times 1440) = 4.7 \times 10^{21}$ cells, which would fill more than 1.35 million soccer balls. (Note: If your answers to this problem are a little different from those given, it is probably because the numbers given were calculated to nine significant digits before rounding off.)

Logistic Population Growth

The calculations in Problem 1.5 indicate that no real population can grow exponentially for more than a relatively small number of generations without catastrophic consequences. In nature, although factors such as disease and predation often contribute to the control of population size, populations that grow too large ultimately must deplete the available resources. The kind of growth curve in Figure 1.7 is typical for populations expanding in a new environment: the initial population growth is exponential, but then the rate of growth gradually decreases.

A simple alternative to exponential growth is the logistic model; the term *logistic* refers to proportions and, in the logistic model, the rate of population growth is assumed to decrease in proportion to the population size. By analogy with Equation 1.4, the change in population size with a discrete model of population growth takes the form

$$N_t = N_{t-1} + rN_{t-1} \frac{(K - N_{t-1})}{K}$$

In this equation, $K$ is a constant known as the carrying capacity of the environment. Observe that, when $N$ is very small compared with $K$, then $N_t = N_{t-1} + rN_{t-1}$, and so population growth is nearly exponential. On the other hand, when $N$ is close to $K$, then $N_t = N_{t-1}$, and so population growth comes to a standstill.

Unlike Equation 1.4, Equation 1.10 does not have a simple solution for $N_t$ in terms of $N_0$. However, if the population grows sufficiently slowly, then population growth can be treated as continuous, and Equation 1.10 yields the differential equation

$$\frac{dN(t)}{dt} = rN(t) \left( \frac{K - N(t)}{K} \right)$$

The solution of Equation 1.11 is given by

$$N(t) = \frac{K}{1 + Ce^{-rt}}$$
where the constant \( C = (K - N_0)/N_0 \). Equation 1.12 is called the logistic growth curve and it is derived in Problem 1.7 below. Logistic population growth results in a sort of S-shaped curve like that shown in Figure 1.7, where the parameters are \( r = 0.5355 \), \( N_0 = 10 \), and \( K = 665 \). (Note that the \( r \) and \( N_0 \) parameters are the same as in the exponential-growth model for the same data.) The fit is obviously very good indeed.

PROBLEM 1.6 Use Equation 1.12 with \( N_0 = 10 \), \( r = 0.5355 \), and \( K = 665 \) to calculate \( N(t) \) for the times \( t = 7 \) and \( t = 13 \) and \( t = 14 \). What are the values of \( r \) in Equation 1.10 for \( t = 8 \) and \( t = 14 \)? Why are they not equal to 0.5355? Why are they not equal to each other?

ANSWER With the given parameters, \( N(7) = 261.53 \), \( N(8) = 349.43 \), \( N(13) = 626.13 \), and \( N(14) = 641.68 \). Solving Equation 1.10 for \( r \) and substituting \( N(t) \) yields \( r = 0.5540 \) for \( t = 8 \) and \( r = 0.425 \) for \( t = 14 \). Neither of these values agrees with \( r = 0.5355 \), nor do they agree with each other, because Equation 1.10 pertains to a discrete model and Equation 1.12 to a continuous model. When the population grows continuously, the value of \( r \) needed to produce a given change in population size in some discrete interval of time differs according to the magnitude of the change in population size.

PROBLEM 1.7 Use the expression \( \int \frac{dx}{x^a + bx} = -\frac{1}{a} \ln (x + bx)/x \) to derive the logistic growth curve from Equation 1.11.

ANSWER Write Equation 1.11 as \( dN(t)/dN(t)(K - N(t)) = rK \), so that, comparing with the integral form, it is clear that \( a = K \) and \( b = -1 \). Integrating both sides in accordance with the formula results in \( -(1/K) \ln (K - N(t)/N(t)) = r/K + c \), where \( c \) is a constant of integration chosen so that \( N(t) = N(0) \) when \( t = 0 \). Hence, \( c = -(1/K) \ln (K - N(0))/N(0) = -(1/K) \ln C \), where \( C \) is the constant appearing in Equation 1.12. Consequently, \( \ln (K - N(t))/N(t) = -rt + C \), and so \( (K - N(t))/N(t) = C \exp -rt \). Equation 1.12 follows after some simplification.

SUMMARY Population genetics is the application of Mendel's laws and other genetic principles to entire populations of organisms. It includes the study of genetic variation within and between species and attempts to understand the processes resulting in adaptive evolutionary changes in species through time. Population genetics has many practical applications in medicine, agriculture, conservation, and other fields.

A gene is a hereditary determinant transmitted from parent to offspring that influences a hereditary trait, often in combination with other genes and also with the environment. Alleles are alternative forms of a gene. Genotypes are formed from pairs of alleles and are either homozygous (if the alleles in the genotype are the same) or heterozygous (if the alleles are different). The physical or biochemical characteristics of an organism constitute its phenotype. The essential mechanism of genetic transmission was established in experiments by Gregor Mendel in the years 1856 to 1863. Mendel showed that the alleles of each gene separate (segregate) from one another in the formation of reproductive cells or gametes. Genes are arranged in linear order along chromosomes. A chromosome may contain several thousand genes. Alleles of different genes present in the same chromosome tend to be inherited together (linkage), but the allele combinations can be broken up by recombination.

Chemically, a gene is a region of a DNA molecule. DNA is a metaphoric “twisted ladder” consisting of two paired strands composed of polymers of nucleotides (the sidepieces of the ladder) whose bases (either A, T, G, or C) run inward from the sidepieces to form the rungs. Each rung of the ladder consists of either an A-T base pair or a G-C base pair. Most genes code the polypeptide chains of proteins through a transcript of RNA that is processed into the messenger RNA (mRNA). The polypeptide is produced stepwise by translation of the mRNA according to a triplet genetic code, in which each nonoverlapping group of three adjacent bases (a codon) specifies the amino acid to be attached to the growing chain. Alleles differ in their sequence of nucleotides. A nucleotide substitution in the third position of a codon may result in an amino replacement in the encoded polypeptide because of redundancy in the genetic code. However, most nucleotide substitutions in either of the first two positions do result in amino acid replacements.

A probability is a number between 0 and 1 that measures the likelihood of a particular event being realized in an actual or conceptual experiment. The addition rule applies to mutually exclusive events and states that the probability of one or the other event being realized equals the sum of the separate probabilities. The multiplication rule applies to independent events and states that the probability of both events being realized simultaneously equals the product of the separate probabilities. The probabilities of various
outcomes of repeated and independent trials can be deduced by application of the addition and multiplication rules and conforms to successive terms in the binomial expansion \((p + q)^n\).

Natural populations contain genetic variation in the form of multiple alleles of many genes. For any specified allele, the allele frequency is the proportion of all alleles of the gene that are of the specified type. The allele frequency in a population must usually be estimated from a sample, and so there is variation in the estimate from one sample to the next. The variation is quantified by the standard error. If the distribution of the estimates conforms to a normal, bell-shaped distribution, then the proportions of the estimates lying within \(\pm 1, \pm 2, \text{ and } \pm 3\) standard deviations of the true value of the parameter are 68%, 95%, and 99.7%, respectively. Estimates are also often presented as a confidence interval, which expresses the degree of confidence that the true value of a parameter lies in some specified interval.

A model is a deliberate simplification of a complex situation. Models may be experimental or conceptual. Conceptual models may be verbal, computational, or mathematical. Mathematical models are widely used in population genetics. They specify the mathematical relations between measured or measurable quantities that determine the changes in allele frequency in populations. Population growth affords an example of mathematical modeling. In the simplest model of discrete population growth, at discrete times a constant fraction of the population reproduces, and so the population jumps instantaneously from one size to the next. A more realistic model envisages continuous reproduction throughout time, in which case population growth is exponential. The exponential model often fits population growth in newly colonized environments when the population density is low. Population growth is ultimately limited by nutrients, space, or other resources. When population growth decreases in proportion to population size, the S-shaped logistic curve of population growth results; this curve is determined by the intrinsic rate of increase \(r\) and the carrying capacity of the environment \(K\).

PROBLEMS

1. If you were to catch a collection of Drosophila, grind each one individually in a buffer solution, and measure the rate at which this crude whole-fly homogenate catalyzed the reaction for glucose-6-phosphate dehydrogenase, you would find that the activities would vary by more than four fold. Make a list of possible causes of this variation.

2. Given the complexity of causes of variation in Problem 1, how much variation would you expect to see in the underlying genetic cause of a human inborn error of metabolism such as phenylketonuria? This disorder is caused by insufficient activity of phenylalanine hydroxylase.

3. There are 64 codons in the genetic code, and each codon can undergo nine single-site mutations (each base can mutate to three other bases), for a total of 576 mutations. How many of these result in no change in the "meaning" of the encoded sequence?

4. Assuming that all nucleotides in all codons mutate with equal frequency (i.e., that all 576 mutations in Problem 3 occur at the same rate), are mutations from one amino acid to another all equally likely?

5. The correspondence between genotype and phenotype is one of the most complex and difficult aspects of evolutionary genetics. Describe an example of a gene whose mutations cause more than one distinctly different phenotype that do not appear to be related.

6. A population cage of Drosophila melanogaster is started with 50 males and 50 females, all having the genotype \((rr bb CC DD)\). This notation implies that one chromosome has the \(r\) and \(b\) mutations, and the other has the wild type allele at both loci. These two loci show a frequency of recombination in females of \(r = 0.37\), and the males produce only non-recombinant gametes. Calculate the expected frequency of the gametes for both males and females and the expected offspring genotype frequencies.

7. In some human cultures it is very important to have a son and a daughter, and couples continue having offspring until they have one of each. If an entire population followed this rule, what would happen to the sex ratio in the population?

8. If two genes are on different chromosomes, the probability that a gamete has a particular allele of each of the two genes is the product of the probability of drawing each allele because the draws are independent of one another (see the multiplication rule). If each gene is on a different chromosome, what is the chance that genotype \(Aa Bb CC Dd\) produces two consecutive gametes that are \(ABCD\)?

9. If individual \(X\) has an autosomal recessive disease and both parents are unaffected, what is the chance that the sibling of \(X\) is a heterozygous carrier?

10. A line of mice seems to consistently produce 55% male and 45% female offspring. In order to test whether this deviation is significant, how many offspring would you have to count to be able to reject a 50-50 sex ratio at a probability of \(\alpha = 0.05\)? (Assume that the sex ratio of the mice remains 55:45.)

11. A species of butterflies occurs in two distinct morphs, A and B. You sample two areas and count 26 A and 28 B butterflies in one area, and 10 A and 21 B in another area. Is it possible that these two samples could come from a single homogeneous population, or are the frequencies of the two morphs significantly different from one another?

12. Levy and Levin (1975) used electrophoresis to study the phosphoglucone isomerase-2 gene in the evening primrose Oenothera biennis, a complex genomic heterozygote made true breeding by chromosomal translocations. They observed two alleles affecting electrophoretic mobility of the
enzyme, and among 57 strains they found 35 PG1-2a/PG1-2a, 19 PG1-2a/PG1-2b, and 3 PG1-2b/PG1-2b genotypes.

a. Calculate the allele frequencies of PG1-2a and PG1-2b.

b. With random mating, what would be the expected numbers?

13 The simple models of population growth fail to take into account many factors that affect rates of change. The global human population at 0 A.D. was estimated in millions of people as 200, 200, 200, 200, 200, 250, 250, 400, 500, 900, and 6000. If the population were growing exponentially, these points would fall on a straight line when plotted on a logarithmic scale. Draw this plot. What do you conclude?

14 A healthy pair of Drosophila can produce 500 offspring in 12 days, each adult fly weighing about 1 mg. Assume that the parental flies die after they finish reproducing. (Actually, they live about a month.) If all successive generations get enough to eat and remain this fecund, what will the mass of flies be in one year?

CHAPTER 2

Genetic and Phenotypic Variation

GENETIC VARIATION IN POPULATIONS became a subject of scientific inquiry in the late nineteenth century prior even to the rediscovery of Mendel's paper in 1900. The leading exponent of the study of hereditary differences among human beings was Francis Galton (1822-1911). Galton was a pioneer in the application of statistics to biology. He used statistical methods to study physical traits such as eye color and fingerprint ridges as well as behavioral traits such as temperament and musical ability. Galton was among the first to examine the statistical relations between the distributions of phenotypic traits in successive generations. He is regarded as the founder of biometry, the application of statistics to biological problems.

PHENOTYPIC VARIATION IN NATURAL POPULATIONS

Galton and Mendel exemplify opposite approaches to the study of inherited traits. Mendel's point of departure was discrete variation, in which phenotypic differences among organisms can be assigned to a small number of clearly distinct classes, such as round versus wrinkled peas. Galton's point of departure was continuous variation, in which the phenotypes of organisms are measured on a quantitative scale, like height or weight, and in which the phenotypes grade imperceptibly from one category into the next. As material for the study of phenotypic variation, Galton's choice was good: most of the differences among normal people that are vis-

...
ible to the unaided eye are differences in continuous traits—height, weight, skin color, hair color, facial features, running speed, shoe size, and so forth. The same is true of phenotypic variation in other organisms. On the other hand, as material for the study of genetic variation, Mendel’s choice was good: the pattern of segregation of alleles is revealed most clearly in pedigrees of discrete, simple Mendelian traits.

**Continuous Variation: The Normal Distribution**

With continuous traits, not only do the phenotypes grade into one another, but the traits also usually present difficulties for genetic analysis. The problems are of two principal types:

- Most continuous traits are influenced by the alleles of two or more genes, hence the segregation of any one gene in pedigrees is obscured by the segregation of other genes that affect the trait.
- Most continuous traits are influenced by environmental factors as well as by genes, and so genetic segregation is obscured by environmental effects.

These problems are not insurmountable in organisms with a sufficiently high density of genetic markers scattered throughout the genome (the complement of chromosomes) because the genetic markers can be tracked in pedigrees along with the continuous trait of interest. Organisms with sufficiently dense genetic maps include human beings, laboratory animals, and many domesticated animals and crop plants.

In Galton’s time, however, studies of continuous traits based on genetic linkage were unknown. Why, then, did Galton focus on continuous traits? Because they have a sort of regularity—a statistical predictability—of their own. For many continuous traits, when the phenotypes are grouped into suitable intervals and plotted as a bar graph, the distribution of phenotypes conforms closely to the normal distribution, the symmetrical, bell-shaped curve discussed briefly in Chapter 1 in the section on phenotypic diversity and genetic variation. For example, a bar graph of Galton’s data on the heights of 1329 men, rounded to the nearest inch, is plotted in Figure 2.1. The smooth curve is the normal distribution that best fits the data. The equation of the normal curve is

\[ f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]

where \( x \) ranges from \(-\infty\) to \(+\infty\), and \( \pi = 3.14159 \) and \( e = 2.71828 \) are constants.

The location of the peak of the distribution along the \( x \) axis is determined by the parameter \( \mu \), which is the mean, or average, of the phenotypic values. The degree to which the phenotypes are clustered around the mean is determined by the parameter \( \sigma^2 \), which is the variance of the distribution. Mathematically, the variance is the average of the squared difference of each phenotypic value from the mean; that is, it is the average of the values of \((x-\mu)^2\). How \( \mu \) and \( \sigma^2 \) are estimated from data is considered next.

**Mean and Variance**

Because \( \mu \) and \( \sigma^2 \) are parameters, their values are unknown, and they must be estimated from the data themselves. The height data are tabulated in Table 2.1, in which \( f(x) \) is the number of men whose height is \( x \), rounded to the nearest inch. (The fact that the shortest and tallest men are grouped in the tails of the distribution makes no difference because these men account for only a small proportion of the total sample.) Also tabulated are the products \( f_i \times x_i \) and \( f_i \times x_i^2 \), as well as their sums.

The mean \( \mu \) of the distribution is estimated as the mean of the sample, which is conventionally denoted \( \bar{x} \) (also sometimes as \( \mu \)).

\[ \bar{x} = \frac{\sum f_i x_i}{\sum f_i} \]

In this example, \( \bar{x} = 91.639/1329 = 68.95 \) inches.

Likewise, the variance \( \sigma^2 \) of the distribution is estimated as the variance of the sample, which is conventionally denoted \( s^2 \) (also sometimes as \( \sigma^2 \)).

\[ s^2 = \frac{\sum f_i (x_i - \bar{x})^2}{\sum f_i} = \frac{\sum f_i x_i^2}{\sum f_i} - (\bar{x})^2 \]
The expression in the middle follows directly from the definition of the variance; it is the average of the squared deviations from the mean because, for each value of \( x \), \((x - \overline{x})^2\) is the deviation of that value from the mean. The expression on the right is identical arithmetically but easier to apply in practice. In the example in Table 2.1, \( s^2 = 6,326,939/1329 - (68.96)^2 = 6.11 \). (This value may differ slightly from your own calculation because of round-off error. If the sample size is small (say, less than 50), then a slightly better estimate of the variance is obtained by multiplying the expression in Equation 2.3 by \( n/(n - 1) \), where \( n \) is the total size of the sample (in this case, 1329).

Closely related to the variance is the standard deviation of the distribution, which is the square root of the variance. The standard deviation is a natural quantity to consider in view of the units of measurement. In Table 2.1, for example, each measurement is in inches. The mean is also in inches. However, the variance, being the average of squared deviations, has the units of squared inches—which seems more appropriate for an area than for a height. Taking the square root of the variance restores the correct unit of measure in this example, inches. The estimate of the standard deviation is conventionally denoted \( s \) (also sometimes as \( \sigma \)) and it is calculated as the square root of the quantity in Equation 2.3. In the height example, \( s = 2.47 \) (which may again differ slightly from your own calculation because of round-off error).

The estimate \( s \) of the standard deviation is often called the standard error. When estimating a proportion—such as the frequency of an allele in a population—the standard error is calculated according to Equation 1.3 in Chapter 1.

In Chapter 1, the values 68%, 95%, and 99.7% quoted as the proportions of observations expected to fall within 1, 2, or 3 standard errors of the mean, respectively, emerge directly from Equation 2.1 for the normal distribution. In a normal distribution, the exact proportion of observations falling with any specified range of \( x \) equals the integral of Equation 2.1 across the specified range. For the normal distribution, the integral between the limits \( \mu \pm \sigma \) equals 0.6827, that between \( \mu \pm 2\sigma \) equals 0.9545, and that between \( \mu \pm 3\sigma \) equals 0.9973. In data analysis, \( \overline{x} \) and \( s \) are used in place of \( \mu \) and \( \sigma \). Incidentally, the integral of the normal distribution between the limits \( \mu \pm 4\sigma \) equals 0.9999; this result says that fewer than one in 10,000 observations falls more than four standard deviations from the mean.

**Central Limit Theorem**

Gallton was immensely impressed with the observation that many natural phenomena follow the normal distribution. He writes:

> I know of scarcely anything so apt to impress the imagination with the wonderful form of cosmic order expressed by the "law of frequency of error" [the normal distribution]. Whenever a large sample of chaotic elements is taken in hand and marshaled in the order of their magnitude, this unexpected and most beautiful form of regularity proves to have been latent all along. The law would have been personified by the Greeks if they had known of it. It reigns with serenity and complete self-effacement amidst the wildest confusion. The larger the mob and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of unreason.

It is, indeed, remarkable to consider that pure, blind chance is the reason for this "unexpected and most beautiful form of regularity."

The theoretical basis of the normal distribution is known in probability theory as the central limit theorem. Roughly speaking, the central limit theorem states that the sum of a large number of independent random quantities always converges to the normal distribution. For our purposes, "independent" in this context means that information about any one of the observations gives no improvement in the ability to predict any other of the observations. A large number of independent random quantities is apparently what Galton meant by "a large sample of chaotic elements." The central limit theorem explains in part why so many continuously distributed traits conform to the normal distribution. Most continuous traits are multifactorial, meaning that they are influenced by "many factors," typically several or many genes acting together with environmental factors. Among human

<table>
<thead>
<tr>
<th>Height interval (i)</th>
<th>Height range (in.)</th>
<th>Nearest inch (x)</th>
<th>Number of men (1)</th>
<th>( i \times x )</th>
<th>( i \times x^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>&lt;63.5</td>
<td>63</td>
<td>73</td>
<td>1,449</td>
<td>91,287</td>
</tr>
<tr>
<td>2</td>
<td>63.5-64.5</td>
<td>64</td>
<td>20</td>
<td>1,280</td>
<td>81,920</td>
</tr>
<tr>
<td>3</td>
<td>64.5-65.5</td>
<td>65</td>
<td>64</td>
<td>4,160</td>
<td>270,400</td>
</tr>
<tr>
<td>4</td>
<td>65.5-66.5</td>
<td>66</td>
<td>110</td>
<td>7,260</td>
<td>479,160</td>
</tr>
<tr>
<td>5</td>
<td>66.5-67.5</td>
<td>67</td>
<td>155</td>
<td>10,385</td>
<td>695,795</td>
</tr>
<tr>
<td>6</td>
<td>67.5-68.5</td>
<td>68</td>
<td>199</td>
<td>13,532</td>
<td>920,176</td>
</tr>
<tr>
<td>7</td>
<td>68.5-69.5</td>
<td>69</td>
<td>203</td>
<td>14,007</td>
<td>966,483</td>
</tr>
<tr>
<td>8</td>
<td>69.5-70.5</td>
<td>70</td>
<td>199</td>
<td>13,860</td>
<td>970,200</td>
</tr>
<tr>
<td>9</td>
<td>70.5-71.5</td>
<td>71</td>
<td>171</td>
<td>12,141</td>
<td>862,011</td>
</tr>
<tr>
<td>10</td>
<td>71.5-72.5</td>
<td>72</td>
<td>88</td>
<td>6,336</td>
<td>456,192</td>
</tr>
<tr>
<td>11</td>
<td>72.5-73.5</td>
<td>73</td>
<td>47</td>
<td>3,431</td>
<td>250,463</td>
</tr>
<tr>
<td>12</td>
<td>73.5-74.5</td>
<td>74</td>
<td>27</td>
<td>1,998</td>
<td>147,852</td>
</tr>
<tr>
<td>13</td>
<td>&gt;74.5</td>
<td>75</td>
<td>24</td>
<td>1,800</td>
<td>135,000</td>
</tr>
</tbody>
</table>

Total: 1329

Source: Data from Gallton 1899.
beings, for example, the obvious differences between normal people in hair color, eye color, skin color, stature, weight, and other such traits are not usually traceable to single genes. They result from the combined effects of several or many genes as well as numerous environmental effects acting together as a large sample of chaotic elements, which often produce, in the aggregate, a normal distribution of phenotypes.

It should be emphasized that the "large number" of random elements specified in the central limit theorem need not be excessive. As an example, Figure 2.2 is a bar graph of 100 observations in which each "observation" consists of the sum of nine consecutive random numbers chosen with equal probability from anywhere in the range (-1, +1). For the sum of nine random numbers in this range, the theoretical mean equals 0 and the theoretical standard deviation equals 1.73; the sample values were \( \bar{x} = -0.12 \) and \( s = 1.70 \). Expressed as a deviation from the mean in multiples of the standard error, the number of observations in each category is shown at the top of the bar in Figure 2.2. Because the expected numbers are 2.5, 13.5, 68, 13.5, and 2.5, the fit to a normal distribution is obviously very good. In this example, therefore, fewer than 10 "chaotic elements," when added together, yields "this unexpected and most beautiful form of regularity."

**Figure 2.2** Distribution of 100 values of the sum of nine random numbers from the interval (-1, +1).

**PROBLEM 2.1** At an International Health Exhibition in London in 1884, Galton set up an "anthropometric laboratory" that carried out tens of thousands of measurements covering a wide range of human traits. Among the traits was "strength of pull," expressed as the number of pounds that one could pull with one arm against a resisting force in a sort of arm-wrestling contraption (Galton 1889). The data for 519 males aged 23-26 years fell into the following categories (the number in parentheses is the number of males in each category): 40-50 lbs (10), 50-60 (42), 60-70 (140), 70-80 (168), 80-90 (113), 90-100 (22), 100-110 (24). Using the midpoint of each category as the strength of pull for all males in that category, estimate the mean and standard deviation of strength of pull. Assuming that strength of pull has a normal distribution with parameters equal to these estimates, what is the expected proportion of males whose strength of pull exceeds 112 pounds?

**ANSWER** The values of \( x \) are 45, 55, 65, and so forth. Then \( \bar{x} = 519 \), \( \sum x = 38,675 \), and \( \sum x^2 = 2,963,375 \). Hence, \( \bar{x} = 74.5 \) lbs, \( s^2 = 156.8 \) lbs².

and so \( s = 12.5 \) lbs. (Answers may differ slightly because of round-off error.) A strength of pull of 112 lbs is three standard errors above the mean; hence a proportion of only \((1 - 0.997)/2 = 0.0015\) (about one in 667) males is expected to have a phenotype exceeding this value.

**Discrete Mendelian Variation**

Discrete Mendelian variation (also called simple Mendelian variation) refers to phenotypic differences resulting from segregation of the alleles of a single gene. Environmental effects on the trait are small enough, relative to hereditary differences, that the transmission of alleles determining the trait can be traced through pedigrees. An example of discrete Mendelian variation is the inheritance of red, pink, or white flower color in snapdragons (Chapter 1). This case is exceptionally convenient for genetic studies because of the intermediate phenotype of the heterozygote. However, most of the phenotypic variation in natural populations is multifactorial. In human beings, for example, although simple Mendelian variation accounts for many inherited disorders, each of the disorders is relatively rare.

Ironically, simple Mendelian variation is more easily detected by studying genes and their products than by studying phenotypes. Because the mechanisms of transcription, RNA processing, and translation are relatively free of the gene interactions and environmental effects that complicate the analysis.
of multifactorial traits at the phenotypic level, there is a direct connection between DNA sequences and alleles and a nearly direct connection between genes and their products. Indeed, the correspondence between DNA sequences and alleles is one-to-one different alleles have different DNA sequences irrespective of whether the alleles affect phenotype. Likewise, alleles with nonsynonymous codon differences in a protein-coding region result in different amino acid sequences irrespective of what the polypeptide does in metabolism or how the difference in sequence affects the organism.

Hence, an efficient way to detect simple Mendelian variation is to study molecules—and therein lies a paradox. As evolutionary biologists, populational geneticists are interested in observable phenotypes that are likely to be subject to natural selection: morphology, rate of development, mating behavior, age of reproduction, longevity, and so forth (in short, the types of traits that attracted Galton). On the other hand, genetic studies are most readily carried out with simple Mendelian variation detected as differences between molecules. The paradox is that differences in molecules among healthy organisms are not usually related in any obvious way to differences in phenotype. Thus, there is a gap in being unable to specify exactly which types of molecular differences underlie the evolutionary process. The irony of the situation is similar to that described by the physiologist Albert Szent-Györgyi:

My own scientific life was a descent from higher to lower dimensions, led by the desire to understand life. I went from animals to cells, from cells to bacteria, from bacteria to molecules, from molecules to electrons. The story had its irony, for molecules and electrons have no life at all. On my way, life ran out between my fingers.

The gap between genotype and phenotype results from the complex interactions between genes and environment in the determination of physiology, development, and behavior. In evolutionary biology, the complexity is even greater because the key issue is the relative ability of organisms to survive and reproduce in their environments. Nevertheless, the disconnect between differences in molecules and evolutionary adaptations is by no means inevitable, permanent, or insurmountable. It is already clear that the study of the relationship between genetic variation and evolutionary adaptation must be high on the agenda of evolutionary biology for the next century, and already there are many examples in which the relation is quite well established.

**EXPERIMENTAL METHODS FOR DETECTING GENETIC VARIATION**

For nearly 50 years, the workhorse method for revealing genetic variation has been electrophoresis because small differences in rate of migration in an electrophoretic field can be used to distinguish between nearly identical macromolecules. A typical laboratory setup for electrophoresis is illustrated schematically in Figure 2.3. The tray contains a thick layer of a gel, typically starch, acrylamide, or agarose; it may be placed horizontally (as shown in the illustration) or vertically (with the gel sandwiched between two glass plates). Each sample of material is placed in a small slot near the edge of the gel. Connected to each edge of the gel is a chamber containing a buffered solution and electrodes. In electrophoresis, an electric current is applied across the gel for several hours. Molecules in the samples—usually proteins or nucleic acids—are of greatest interest—move through the gel in response to the electric field. Molecules of different size and charge move at different rates. After the electrophoresis is finished, the positions of the molecule or molecules of interest are revealed by any of several procedures.

**Protein Electrophoresis**

In protein electrophoresis, used primarily to study enzyme molecules, the position to which a particular enzyme migrates is revealed by soaking the gel in a solution containing a substrate for the enzyme along with a dye that precipitates where the enzyme-catalyzed reaction takes place. A dark band thus appears in the gel at the position of the enzyme. If the enzyme present in a sample has an amino acid replacement that results in a difference in the overall ionic charge of the molecule, then the enzyme will have a somewhat altered electrophoretic mobility and move at a different rate. The electrophoretic mobility changes because enzymes of the same size and shape move at a rate determined largely by the ratio of the number of positively charged amino acids (primarily lysine, arginine, and histidine) to the num-
ber of negatively charged ones (principally aspartic acid and glutamic acid). Electrophoresis can therefore be used to detect a mutation that results in a difference in electrophoretic mobility of the enzyme it encodes.

One possible result of an electrophoresis experiment is shown in the hypothetical gel in Figure 2.4A, in which all samples manifest an enzyme with the same electrophoretic mobility. The result indicates a monomorphic sample because there is only one electrophoretic pattern observed. Another kind of result is shown in Figure 2.4B, in which polymorphism is observed in the types of electrophoretic patterns. When polymorphic enzyme bands are observed, genetic tests typically indicate that organisms with only a fast-migrating enzyme are homozygous for a fast allele (F/F) and those with only a slow-migrating enzyme are homozygous for a slow allele (S/S). Organisms with both enzyme bands are heterozygous for the alleles (F/S). Simple Mendelian inheritance of the polymorphism is indicated by, for example, the finding that matings of two heterozygotes produce, on the average, 1/4 F/F, 1/2 F/S, and 1/4 S/S progeny. Two enzyme bands appear in heterozygotes whenever the active enzyme consists of a single polypeptide chain (rather than two or more polypeptide chains aggregated together) because heterozygotes produce a different polypeptide chain from each allele.

Enzymes that differ in electrophoretic mobility as a result of allelic differences in a single gene are called allozymes. Hence, allozyme variation in a population is an indication of simple Mendelian genetic variation. Allozyme variation is widespread in almost all natural populations studied by electrophoresis, including organisms such as bacteria, plants, Drosophila, mice, and human beings.

**PROBLEM 2.2** A sample of 35 organisms from a Texas population of the wild annual plant *Pilosum drumondii* were examined for the electrophoretic mobility of the enzyme alcohol dehydrogenase (Levin 1978). Two alleles affecting electrophoretic mobility were found—Adh* and Adh* The genotype frequencies observed in the sample were 0.04 Adh*/Adh*, 0.32 Adh*/Adh*, and 0.64 Adh*/Adh*. Estimate the allele frequency of Adh* and its standard error.

**ANSWER** Let \( p \) represent the allele frequency of Adh*. Then \( \hat{p} = 0.04 + 0.32/2 = 0.20 \). The standard error equals \( \sqrt{(0.20)(1-0.20)/35} = 0.05 \).

**PROBLEM 2.3** From a natural population of *Drosophila melanogaster* in Raleigh, North Carolina, 660 fertilized females were trapped and used to found a large laboratory population (Mukai et al. 1974). After about five months (10 generations), 489 third chromosomes in the population were examined for allozymes coding for the enzymes enolase-6 (alleles Esh and Esh*), esterase-C (alleles EC and EC*), and octanol dehydrogenase (alleles Odh and Odh*). The order of the genes in the third chromosome is known to be Esh-EC-Odh. The results were as follows:

<table>
<thead>
<tr>
<th>Esh</th>
<th>EC</th>
<th>Odh</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>35</td>
<td>64</td>
<td>76</td>
<td>152</td>
</tr>
<tr>
<td>7</td>
<td>64</td>
<td>74</td>
<td>13</td>
</tr>
<tr>
<td>15</td>
<td>64</td>
<td>1</td>
<td>29</td>
</tr>
<tr>
<td>1</td>
<td>66</td>
<td>0</td>
<td>8</td>
</tr>
</tbody>
</table>

Estimate the allele frequencies and their standard errors for Esh and Esh*, for EC and EC*, and for Odh* and Odh*. What number of each of the chromosome types is expected assuming that the alleles are associated at random?