CONTENTS

1

Mathematical and Biological Introduction

1.1 Interpreting Population-Genetic Statistics

Patterns of genetic variation in populations reflect the outcome of the evolutionary processes that have shaped the populations in the past. Many types of evolutionary phenomena, including the classic effects of gene flow, mutation, natural selection, and changes in population size, have distinctive consequences for patterns of genetic variation. Much of the field of population genetics—the branch of evolutionary biology concerned with genetic variation in populations—involves the measurement of genetic variation, applying the long tradition of population-genetic theory to understand how evolutionary mechanisms generate patterns in genetic variation, and connecting observations of genetic variation to their consequences for the underlying biological phenomena.

This enterprise necessarily relies on *statistics*—that is, functions computed from data—to evaluate the features of genetic variation. In a schematic paradigm for population-genetic data analysis, an investigator first begins with an interest in empirical biological phenomena in population genetics and related fields (Figure 1.1). The investigator collects measurements of variable genetic types in populations, with the aim of understanding these phenomena. Summary statistics—functions that provide summaries of complex data sets—are then applied to the genetic variation data. To interpret the values of the summary statistics, investigators can draw on theoretical modeling, simulations of evolutionary models, and past empirical studies that connect particular patterns in the summary statistics to aspects of the biological phenomena of interest.

Our focus here is on another important factor that influences the interpretation of population-genetics summary statistics: the mathematical properties of the summary statistics themselves. One of the most important classes of statistics utilized in population-genetic studies is the set of statistics that are computed from *allele frequencies*, quantities that represent the frequencies with which particular genetic types occur in a population. Allele frequencies and the statistics that employ them are among the oldest, most fundamental, and most widely applied quantities in population genetics. Even as population-genetic data sets have grown dramatically in size in the era of genome sequencing, and even as statistical methods have advanced in sophistication and computational complexity, summary statistics have continued to provide the foundation for population-genetic

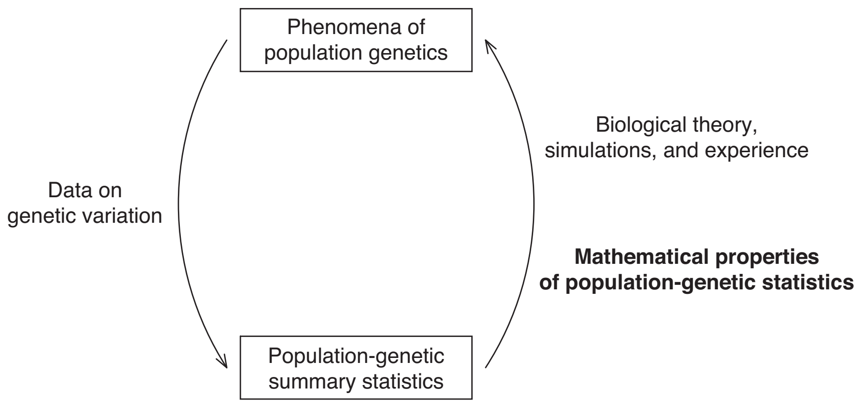

FIGURE 1.1. A conceptual model for population genetics, emphasizing the role of the mathematical properties of population-genetic statistics in inference about population-genetic phenomena from data on genetic variation.

data analysis, serving as a guide to developing intuition about the biological phenomena that underlie the data.

Despite their ubiquity throughout the long history of the field, simple summary statistics can have surprising mathematical properties, properties that can have a substantial impact on the way in which the statistics are interpreted. The allele frequencies at a genetic site represent a collection of nonnegative numbers that sum to 1. This simple fact generates endless consequences for the statistics commonly used for analyzing genetic variation in populations, and unexpected phenomena observed in patterns of genetic variation can often be traced to underlying mathematical features of collections of nonnegative numbers whose sum is 1. In particular, statistics can have upper and lower bounds that depend on aspects of the allele frequencies or on close mathematical relationships to other such statistics.

As we will see, the insights provided by a careful mathematical treatment of population-genetic statistics enable us to explain a number of otherwise surprising observations that have been made from population-genetic data. We will examine how counterintuitive features of population-genetic summary statistics can be produced by the often-unexpected or underappreciated phenomenon that the mathematical upper and lower bounds on the statistics can vary with aspects of the allele frequencies. To reduce the potential for misinterpretations, we will study the bounds on a variety of population-genetic statistics in relation to other such statistics, with a focus on measures of genetic homogeneity and diversity built from the concept of homozygosity. These bounds can facilitate sensible biological understanding, and in some instances they can suggest useful normalized statistics. We examine how the mathematical bounds on population-genetic statistics provide insight into potentially counterintuitive phenomena observed in such contexts as testing for deviation from population-genetic null models and uncovering signatures of natural selection. The aim is to thoroughly investigate how statistics used in population genetics depend on allele frequencies, and to provide mathematics that informs the use and interpretation of these statistics.

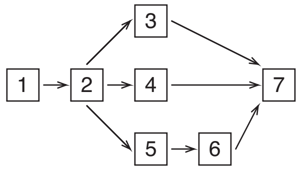

FIGURE 1.2. Dependencies among the chapters in the book.

1.2 A Note about Assumed Background

The book lies at an intersection of mathematics, population genetics, and statistics. As the motivating examples and interpretations of the mathematical results lie in the field of population genetics, the typical reader is expected to have some familiarity with population-genetic concepts and applications, at the level of textbooks in the field. Biologically oriented readers are expected to be aware of typical empirical uses of population-genetic statistics, and of their conventional interpretations. For biological readers interested mostly in the implications of the mathematical results, the mathematics is at an elementary level, primarily utilizing calculus and elementary probability, and only light knowledge of statistics is required.

At the same time, though the motivation derives from biology, many of the results in this book can be stated purely as mathematical results concerning nonnegative numbers that sum to 1—without reference to biology at all. Population-genetic statistics are often analogous or even mathematically identical to corresponding quantities in other areas, including ecology, economics, and the field of statistics more generally. Thus, a number of the mathematical results about population-genetic statistics can be viewed as providing information about these other quantities as well. Although the language of population genetics is used to contextualize the mathematical results, care is taken to separate results that are purely mathematical from their consequences in biology. The work thus offers a basis for readers from other fields that employ the same statistics to understand and make use of the results without requiring extensive knowledge of the biological context.

A diagram of the dependencies among the chapters of the book appears in Figure 1.2. For introductions to population genetics with a focus primarily on biological phenomena, readers are encouraged to consult [29], [52], [57], [60], [64], and [120]. For more emphasis on statistical methods, see [12], [92], and [179].

1.3 Definitions

We focus on a genetic *locus* or *marker*, a location in a genome at which different genetic types can be measured. These distinct types are termed *alleles* or *allelic types*. All that is required of a locus is that it be a genomic region that in principle is assayable in different individuals, and that the result of the assay be classifiable into one of a number of allelic types. The terms *locus* and *marker* are used interchangeably.

We can treat as a locus a single genomic site or *base pair* that can have type A in some genomes and type C, G, or T in others (Figure 1.3A). Alternatively, a locus can be a contiguous piece of DNA sequence consisting of many base pairs (Figure 1.3B). We also

(A)	(B)	(C)
ACGCT	TACGC	TCAGCGATAGATAGATAGGCT
ACACT	CACGC	TCAGCGATAGATAGATAGATAGGCT
ACTCT	TACAT	TCAGCGATAGATAGATAGATAGATAGGCT

FIGURE 1.3. Three loci, each with three allelic types illustrated. In each locus, allelic types are described by a distinct aspect of the DNA sequence. (A) A locus defined by a single genomic site—the third one in the sequence. (B) A locus defined by a contiguous piece of DNA sequence five base pairs long. Distinct sequences over the entire region are distinct allelic types, or haplotypes. (C) A locus defined by a DNA sequence region in which distinct sequences vary in their number of copies of a short repeated segment, GATA. Such a locus is a *microsatellite* or *short tandem repeat* locus.

consider loci termed *microsatellites* and characterized by *short tandem repeats*—regions in which individuals differ in their number of copies of a short repeated segment (Figure 1.3C). Such loci can be regarded as examples of *insertion* or *deletion* loci, in which allelic types at a locus entail differences in presence and absence of specified pieces of DNA.

The number of distinct alleles at a locus in a population is denoted by I; except where otherwise specified, we treat I as finite, but possibly very large. In a *haploid* species, each individual possesses one copy of each genetic locus, so that only one observation of the locus can be obtained from an individual; this observation has allelic type in the set $\{1, 2, \ldots, I\}$. Individuals in a *diploid* species each have two copies, so that two observations can be obtained from an individual. A *polyploid* species has more than two copies of a locus; the *ploidy* of such a species specifies the number of copies.

The *genotype* of a haploid individual at a locus is the single allelic type observed at the locus in the individual. For a diploid individual, the genotype is the pair of allelic types, not considering their order, except where otherwise noted. Thus, for alleles i, j with $i \neq j$, genotypes ij and ji have the same meaning. If a diploid genotype has two identical allelic types, the individual is a *homozygote*. Otherwise, the individual is a *heterozygote*. The (unordered) genotype at a locus of a polyploid individual of ploidy α is the set of all α allelic types it has at the locus.

Note that an observation of the allelic type at a locus is, like the type itself, known as an *allele*. We will often avoid this abuse of terminology by distinguishing between the *allelic types* or distinct alleles at a locus and *observed* or *sampled alleles*; for example, the allelic types at a locus might be A, B, and C, and an individual might be described as having allele A. The meaning of *allele* as a type or as an observation will be clear from the context.

In considering a locus that consists of multiple neighboring sites (Figure 1.3B), an allelic type in a haploid genome is termed a *haplotype*. A haplotype is simply a special case of a genotype for loci consisting of multiple sites; an observation at multiple sites can be classified as having one of a number of distinct haplotypes from the set $\{1, 2, \ldots, I\}$. Alternatively, haplotypes can be classified by the vector of allelic types present across all the sites in the locus. At a locus that consists of multiple sites, an individual with a diploid genome possesses two haplotypes, one for each of the two copies of the genome.

The term *haplotype* is subject to the same abuse of terminology as *allele*, referring both to an observation at a locus defined by multiple sites in a genomic region and to the classification of that observation. We can use *haplotypic types* or distinct haplotypes for the

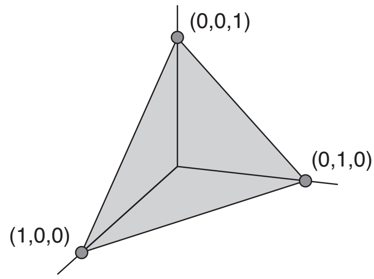

FIGURE 1.4. The unit simplex Δ_2, representing the possible allele frequency vectors for a locus with $I = 3$ distinct alleles.

classification and *observed* or *sampled haplotypes* for observations. As a haplotypic type is a special case of an allelic type for a certain class of loci and an observed haplotype is a special case of an observed allele, in mathematical results about allelic types or observed alleles, it is understood that they also apply to haplotypic types or observed haplotypes.

In a population, the probability that a random observation of a locus has a particular allelic type is the *allele frequency*. The sum across all allelic types at a locus of the allele frequencies in a population is 1. In other words, denoting by p_i the frequency of allelic type i, each p_i lies in $[0, 1]$, and

$$\sum_{i=1}^{I} p_i = 1. \tag{1.1}$$

We refer to the vector of allele frequencies $\mathbf{p}$ at a locus in a population as an *allele frequency distribution* or *allele frequency vector*. We allow alleles to have zero frequency except where otherwise stated. A locus in which one allelic type has frequency 1 and all others have frequency 0 is said to be *monomorphic*, and the lone allelic type is said to be *fixed* in the population. A locus for which at least two allelic types have positive frequency is *polymorphic*.

For $I \geqslant 1$, the vector of allele frequencies at a locus is a point in the unit $(I-1)$-*simplex*, Δ_{I-1}, defined by

$$\Delta_{I-1} = \left\{ (p_1, p_2, \ldots, p_I) \in \mathbb{R}^I \,\middle|\, \sum_{i=1}^{I} p_i = 1 \text{ and } p_i \geqslant 0 \text{ for all } i \right\}. \tag{1.2}$$

The $(I-1)$-simplex Δ_{I-1} represents the set of possible allele frequency vectors with I distinct alleles (Figure 1.4). The vertex of the simplex at which $p_i = 1$ and $p_j = 0$ for $j \neq i$ and allele i is fixed is denoted by $\mathbf{e}_i$.

At this point it is helpful to clarify that in population-genetic studies, a "population" generally refers to a group of organisms that share a feature of interest to the investigator, such as a shared ancestry or shared habitation of a geographic location. Here, we are considering allele frequencies of an extant population in the present; the allele frequencies represent a snapshot of genetic variation in the present, and our mathematical results do not consider the genealogy that underlies them. Note that the field of statistics has a different meaning for "population" as a space of possible observations from which some sample of observations is drawn. Our use of the term "population" follows the usage in population genetics; conveniently, however, the biological populations of interest here are also statistical populations. Except where otherwise specified, the allele frequencies are treated as *parametric*—the true frequencies in the biological population—rather than as estimates obtained from data. As parametric frequencies, they describe a statistical population identified with the biological population. This choice places our analysis in the realm of mathematics rather than in a context of statistical sampling. Alternatively, one can view the allele frequencies as estimates obtained in samples from an infinite (biological) population. In this context, the strong law of large numbers [153] allows us to view sample allele frequencies estimated in the infinite (biological) population as parametric allele frequencies.

We define by Δ the set of all possible allele frequency distributions, $\Delta = \cup_{I=1}^{\infty} \Delta_{I-1}$. The set Δ represents the set of allele frequency distributions when the number of distinct alleles, I, is left unspecified. As loci represented by $\mathbf{e}_i$ are monomorphic, we will also have occasion to consider $\Delta^* = \Delta \setminus \{\mathbf{e}_1, \mathbf{e}_2, \ldots\}$, the set of allele frequency distributions representing polymorphic loci, and $\Delta^*_{I-1} = \Delta_{I-1} \setminus \{\mathbf{e}_1, \mathbf{e}_2, \ldots \mathbf{e}_I\}$, the set of allele frequency distributions representing polymorphic loci with at most I distinct alleles.

For a population of diploids, each distinct genotype ij has a *genotype frequency*, representing the probability that a randomly drawn individual in the population has the genotype. The genotype frequencies at a locus in a diploid population are said to satisfy *Hardy-Weinberg proportions* if (1) for each i, the probability that an individual has genotype ii is p_i^2, and (2) for each unordered i, j with $j \neq i$, the probability that an individual has genotype ij is $2p_ip_j$. If the two observations in a randomly chosen individual from the population represent independent random draws from the allele frequency distribution at a locus, then the genotype frequencies at the locus will satisfy Hardy-Weinberg proportions. Figure 1.5 illustrates a geometric interpretation of Hardy-Weinberg proportions. Note that the term *Hardy-Weinberg proportions* describes the equations satisfied by genotype probabilities. The related term *Hardy-Weinberg equilibrium* refers to the equilibrium point of a dynamical system; at that equilibrium, Hardy-Weinberg proportions are satisfied.

Each statistic that we consider can be written as a function of the allele frequency vector in one population, or as a function of the allele frequency vectors in two or more populations. For example, for a single population with I distinct alleles at a locus, we can define the *frequency of the most frequent allele* as a function $M(p_1, p_2, \ldots, p_I)$,

$$M(p_1, p_2, \ldots, p_I) = \max_{i \in \{1,2,\ldots,I\}} p_i. \tag{1.3}$$

We generally drop the argument $(p_1, p_2, \ldots, p_I)$ for this and other statistics, simply writing M for the frequency of the most frequent allele. Considering all possible allele frequency vectors $\mathbf{p} \in \Delta_{I-1}$, M lies in $(0, 1]$. Like all allele frequencies, M is bounded above by 1. Also, because the sum of the allele frequencies is 1, at least one allele frequency must be positive—in particular, the frequency of the most frequent allele. Note that M is well defined not only on Δ_{I-1}, but also on Δ.

We focus on a detailed study of statistics of *genetic diversity* and *genetic homogeneity*; for completeness, we also define statistics of *genetic differentiation* and *genetic similarity*, as the

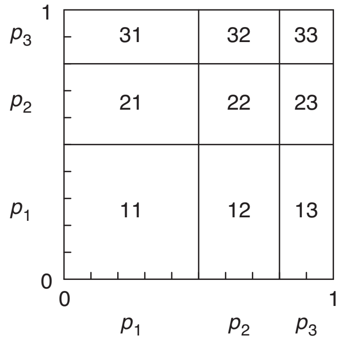

FIGURE 1.5. Hardy-Weinberg proportions for a locus with allele frequencies $p_1 = 0.5$, $p_2 = 0.3$, and $p_3 = 0.2$. Separately on the x- and y-axes, the unit interval is partitioned into three components, representing the frequencies of the three distinct alleles. In each area within the unit square is an ordered genotype, representing a pair of alleles drawn with replacement from the allele frequency distribution. Under Hardy-Weinberg proportions, the frequency of an ordered diploid genotype is the product of the frequencies of its constituent alleles.

general issues of mathematical bounds on population-genetic statistics have often been raised in the context of such statistics, to which we return in Chapter 7. For each statistic, we use the same notation throughout the book, sometimes slightly abusing notation by employing the associated symbol both for the function and for the value of the function computed from a particular allele frequency distribution; whether the function or the value is meant will be clear from the context. To maintain consistent notation throughout the book, we will denote some statistics by letters that differ from those often used in the literature.

We first need a series of mathematical results.

1.4 Standard Inequalities

Our analyses of upper and lower bounds on population-genetic statistics repeatedly employ a number of mathematical results. We state a few standard inequalities. Proofs and additional information about these inequalities can be found in [17], [25], [104], and [160].

1.4.1 Cauchy-Schwarz Inequality

Theorem 1.1: Consider two sequences of nonnegative real numbers, $\{p_i\}_{i=1}^{\infty}$ and $\{q_i\}_{i=1}^{\infty}$. Then

$$\left(\sum_{i=1}^{\infty} p_i^2\right)\left(\sum_{i=1}^{\infty} q_i^2\right) \geqslant \left(\sum_{i=1}^{\infty} p_i q_i\right)^2. \tag{1.4}$$

Equality holds in eq. 1.4 if and only if for all i, $p_i = \lambda q_i$ for a constant λ.

The Cauchy-Schwarz inequality states that the square of the sum of element-wise products of two sequences is bounded above by the product of the separate sums of squares of the two sequences. Note that by setting all except the first I elements of $\{p_i\}_{i=1}^{\infty}$ and $\{q_i\}_{i=1}^{\infty}$ to 0, the inequality holds for sequences of any finite length I.

We will have multiple occasions to use the following corollary of the Cauchy-Schwarz inequality.

Corollary 1.2: Consider a sequence of nonnegative real numbers with length I, $\{p_i\}_{i=1}^{I}$. Define $C = \sum_{i=1}^{I} p_i$. Then

$$\sum_{i=1}^{I} p_i^2 \geqslant \frac{C^2}{I}, \tag{1.5}$$

with equality if and only if $p_1 = p_2 = \ldots = p_I = C/I$.

Proof: Consider sequences $\{p_i\}_{i=1}^{I}$ and $(1, 1, \ldots, 1)$, both of length I. By the Cauchy-Schwarz inequality, $(\sum_{i=1}^{I} p_i^2)(\sum_{i=1}^{I} 1) \geqslant (\sum_{i=1}^{I} p_i)^2$, with equality if and only if $p_i = \lambda$ for a constant λ. Because $\sum_{i=1}^{I} p_i = C$, equality holds if and only if $p_i = C/I$ for all i. □

1.4.2 Rearrangement Inequality

Theorem 1.3: Consider two sequences of real numbers, $\{p_i\}_{i=1}^{I}$ and $\{q_i\}_{i=1}^{I}$, ordered such that $p_1 \geqslant p_2 \geqslant \ldots \geqslant p_I$ and $q_1 \geqslant q_2 \geqslant \ldots \geqslant q_I$. Suppose σ is a permutation of $(1, 2, \ldots, I)$, mapping i to $\sigma(i)$ for $i = 1, 2, \ldots, I$. Then

$$\sum_{i=1}^{I} p_i q_i \geqslant \sum_{i=1}^{I} p_i q_{\sigma(i)} \geqslant \sum_{i=1}^{I} p_i q_{I-i+1}.$$

The rearrangement inequality states that the sum of element-wise products of two sequences, pairing each entry in one of the sequences with an entry from the other sequence, and allowing one of the sequences to be permuted, is maximal when the two sequences are placed in the same order, descending from the greatest value to the smallest. The sum is minimal when one of the sequences is reversed. Although our focus on statistics in a single population means that we will not make use of the rearrangement inequality here, it is included as it is useful for consideration of properties of statistics involving pairs of populations.

1.5 Genetic Diversity and Genetic Homogeneity

Genetic diversity statistics, treated as functions of the allele frequencies of a locus in a single population, measure the level of variability of the locus in the population. A sensible diversity statistic assigns the minimal diversity score to a locus that is not variable, and it has higher values when the locus has multiple allelic types with nontrivial frequencies. A statistic of genetic homogeneity can be easily obtained from a genetic diversity statistic by transforming the values for the diversity statistic so that the largest value occurs for a monomorphic locus and the smallest values occur for multiallelic loci with many nontrivial allele frequencies. It is often convenient to obtain results on statistics of homogeneity,

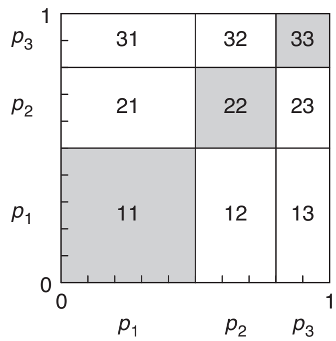

FIGURE 1.6. Homozygosity for a locus with allele frequencies $p_1 = 0.5$, $p_2 = 0.3$, and $p_3 = 0.2$. Separately on the x- and y-axes, the unit interval is partitioned into three components, representing the frequencies of the three distinct alleles. As in Figure 1.5, each area within the square shows an ordered diploid genotype. The homozygosity, represented by the shaded areas, is $J = 0.5^2 + 0.3^2 + 0.2^2 = 0.38$.

and to then perform a transformation in order to obtain results concerning diversity statistics; with this understanding, we will primarily report results in terms of homogeneity statistics.

1.5.1 Heterozygosity and Homozygosity

The *homozygosity* of a locus with allele frequencies $\{p_i\}_{i=1}^{I}$ is a statistic of genetic homogeneity, defined as

$$J = \sum_{i=1}^{I} p_i^2. \tag{1.6}$$

Here, we drop the implied argument $(p_1, p_2, \ldots, p_I)$ for J. The probability that two independent samples drawn from the population produce allelic type i is p_i^2. Hence, considering all allelic types, the homozygosity represents the probability that two independent draws from the population produce the same allelic type.

Considering all possible allele frequency distributions in the simplex Δ_{I-1}, homozygosity J lies in $(0, 1]$. As a sum of squares, with at least one positive term in the sum (M^2), homozygosity of a locus in Δ_{I-1} is positive and bounded below by a positive value. Indeed, we can use Corollary 1.2 to provide the lower bound on J over the set of allele frequency distributions Δ_{I-1} (Exercise 1.1). Figure 1.6 provides a geometric visualization of homozygosity.

A homozygosity of 1 is achieved if $p_i = 1$ for some i, and $p_j = 0$ for all $j \neq i$. In fact, $J = 1$ if and only if a locus is monomorphic. To obtain this result, note that by eq. 1.1,

$$\begin{aligned} J &= \left(\sum_{i=1}^{I} p_i\right)^2 - 2\sum_{i=1}^{I-1}\sum_{j=i+1}^{I} p_i p_j \\ &= 1 - 2\sum_{i=1}^{I-1}\sum_{j=i+1}^{I} p_i p_j. \end{aligned} \tag{1.7}$$

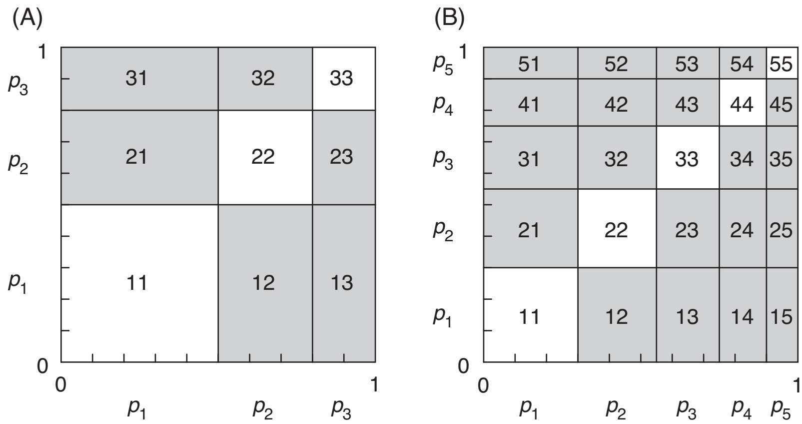

FIGURE 1.7. Heterozygosity for two loci, illustrating a smaller heterozygosity (A) and a larger heterozygosity (B). Separately on the x- and y-axes, the unit interval is partitioned into components representing the frequencies of the different alleles. The heterozygosities, represented by shaded areas are (A) $H = 1 - (0.5^2 + 0.3^2 + 0.2^2) = 0.62$ and (B) $H = 1 - (0.3^2 + 0.25^2 + 0.2^2 + 0.15^2 + 0.1^2) = 0.775$. The locus in (B) has more alleles than the locus in (A), and its allele frequencies are more similar to each other than those in (A); thus, it is sensible to view its allele frequency distribution as "more diverse."

Then $J = 1$ if and only if $2\sum_{i=1}^{I-1}\sum_{j=i+1}^{I} p_i p_j = 0$, which, in turn, occurs if and only if $p_i > 0$ for only one value of i.

To obtain a corresponding statistic that measures genetic diversity, we can consider the *heterozygosity* of a locus $H = 1 - J$, or

$$H = 1 - \sum_{i=1}^{I} p_i^2. \tag{1.8}$$

This quantity represents the probability that two independent draws from the population produce distinct allelic types. Heterozygosity H lies in $[0, 1)$. In the literature, heterozygosity and homozygosity have both been denoted by H; here we consistently use J for varieties of homozygosity and H for heterozygosities. Figure 1.7 illustrates how heterozygosity is a sensible representation of diversity.

The interpretation of homozygosity and heterozygosity in the case of diploid organisms explains the meaning of the terms. For a diploid, homozygosity is the fraction of individuals in the population expected to have two identical copies at the locus, under the assumption of Hardy-Weinberg proportions—that is, the fraction expected to be homozygotes. Similarly, heterozygosity is the fraction of individuals in the population expected to be heterozygotes. Indeed, J and H are often termed *expected homozygosity* and *expected heterozygosity*, respectively, as distinguished from *observed homozygosity* and *observed heterozygosity*. The distinction is important when tabulating the occurrence of homozygotes and heterozygotes in diploid genotypes, as the fractions of homozygotes and heterozygotes so obtained are termed *observed* homozygosities and heterozygosities. The "expected" in the terms "expected homozygosity" and "expected heterozygosity" refers to the expectation of the indicator random variables that record if a randomly drawn individual is a homozygote or a heterozygote under the assumption that two alleles are sampled independently from an allele frequency distribution. We calculate homozygosities and heterozygosities *only* from allele frequency vectors, so that the "observed" statistics are not considered; hence, we omit "expected" in describing homozygosity and heterozygosity.

The quantities termed homozygosity and heterozygosity in population genetics achieved widespread use as statistics with the increasing availability of population-genetic data that accompanied the measurement of protein variation in the 1960s [59, 74, 99], one of the founding papers on protein variation [74] having asked, "At what proportion of his loci can we expect a diploid individual to be heterozygous?" The idea of heterozygosity as a statistic appears in one of the most extensive data analyses of the time [97]; and homozygosity and heterozygosity statistics are prominent in the work of Nei [115, 116, 117, 118, 119]—who has also used the terms *gene identity* and *gene diversity* to refer to homozygosity and heterozygosity, respectively—and in early statistical tests of population-genetic models to detect unusual patterns of genetic variation [177, 178]. As measurements of the probabilities of identity and difference for pairs of alleles in population-genetic models, the concepts of homozygosity and heterozygosity are much older; they trace to early population-genetic theory, where they were sometimes studied as *homozygosis* and *heterozygosis* [185], and terms such as homozygosis, homozygosity, heterozygosis, and heterozygosity are common in early papers in population genetics [43]. Current uses of homozygosity and heterozygosity make use of these concepts in the context of phenomena such as conservation, gene flow, hybridization, inbreeding, natural selection, and relatedness [8].

As has sometimes been noted in the population-genetic context [177], homozygosity and heterozygosity have appeared in a variety of other settings in which objects are classified by type and a measurement of similarity or diversity is of interest. The Herfindahl-Hirschman index in economics [67, 71, 72] is equivalent to homozygosity, except that p_i is interpreted as the market share of firm i, expressed as a proportion of the total size of the market. Simpson's index in ecology [155] is also equivalent to homozygosity; p_i represents the fraction of a collection of individuals that originate from species i, and Simpson's index is the probability that two individuals drawn with replacement from the collection are taken from the same species. The Bice-Boxerman index of concentration of health care in health services research [20] is a homozygosity statistic. Quantities equivalent to heterozygosity have appeared associated with the names Gini or Gini-Simpson, Gibbs-Martin [51], and Blau [22], and in the probability-of-interspecific-encounter statistic in ecology [77].

1.5.2 Extensions to Homozygosity

We also examine a variety of quantities obtained by modifications of the formula for homozygosity. The α-homozygosity changes the exponent to which each allele frequency is raised:

$$J^{(\alpha)} = \sum_{i=1}^{I} p_i^{\alpha}, \tag{1.9}$$

where α is a constant with $\alpha > 1$. The case of $\alpha = 2$ is the standard 2-homozygosity.

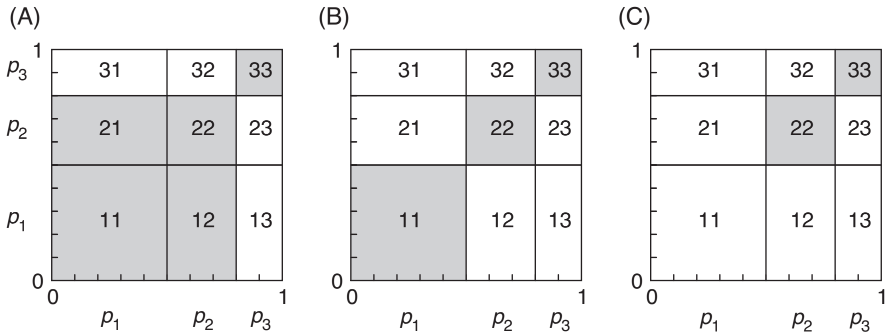

FIGURE 1.8. The quantities J_A, J_B, and J_C for a locus with allele frequencies $p_1 = 0.5$, $p_2 = 0.3$, and $p_3 = 0.2$. Separately on the x- and y-axes, the unit interval is partitioned into three components, representing the frequencies of the three distinct alleles. The shaded areas in the three panels represent J_A, J_B, and J_C, respectively. (A) $J_A = 0.68$. (B) $J_B = 0.38$. (C) $J_C = 0.13$.

We often order the allele frequencies $\{p_i\}_{i=1}^{I}$ so that $p_i \geqslant p_j$ for $i < j$. With this arrangement, M is equivalent to p_1. M can be interpreted as a statistic of genetic homogeneity, since $M = 1$ for a monomorphic locus, and M has small values for a polymorphic locus with many nonzero allele frequencies. With ordered allele frequencies, we can consider additional modifications to homozygosity. Arrange the allele frequencies at a locus in descending order of frequency. Define

$$J_A = (p_1 + p_2)^2 + \sum_{i=3}^{I} p_i^2, \tag{1.10}$$

$$J_B = p_1^2 + p_2^2 + \sum_{i=3}^{I} p_i^2, \tag{1.11}$$

$$J_C = p_2^2 + \sum_{i=3}^{I} p_i^2. \tag{1.12}$$

These statistics arise in a context in which it is of interest to modify the role of the most frequent allele in calculating a homogeneity or diversity statistic. As defined in eq. 1.11, $J_B = J$. The quantity J_A, which we term the 1, 2-*pooled homozygosity*, can be viewed as combining the frequencies of the two most frequent alleles into a single allelic class with frequency $p_1 + p_2$. J_C, the 1-*truncated homozygosity*, drops the contribution of the most frequent allele in computing homozygosity. For any allele frequency vector $\mathbf{p} \in \Delta_{I-1}$, the statistics satisfy $J_A \geqslant J_B > J_C$ (Exercise 1.2).

Considering all allele frequency vectors, J_A and J_B lie in $(0, 1]$, and J_C lies in $[0, 1]$ (Exercise 1.3). Figure 1.8 provides a visualization of all three statistics. J_A, J_B, and J_C refer to the quantities denoted H_{12}, H_1, and H_2 by Garud et al. [48].

Garud et al. [48] also examined

$$Z = \frac{J_C}{J_B}, \tag{1.13}$$

representing the fraction of the homozygosity J_B due to homozygotes for alleles other than the most frequent allele. The quantity Z lies in $[0, 1)$.

1.6 Genetic Differentiation and Genetic Similarity

Statistics of genetic differentiation and genetic similarity are functions of the allele frequency vectors in two or more populations, and they measure the level of difference between the allele frequencies of the different populations. Differentiation statistics have low values when the various populations have identical allele frequency vectors, and high values when the populations have substantially different allele frequency vectors. A statistic of genetic similarity can be obtained by transforming a genetic differentiation statistic so that the largest value occurs when populations have identical allele frequency vectors and the smallest value occurs when populations have substantially different allele frequency vectors.

1.6.1 *Many Populations*

To examine genetic differentiation, we consider a locus in a set of K populations (sometimes viewed as "subpopulations" of a larger population). We assume that in the collection of populations, the locus is polymorphic. Denote the frequency of allele i in population k by p_{ki}. Each population k has an allele frequency vector $\mathbf{p}_k$, and the list of allele frequency vectors in all K populations, $(\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_K)$, lies in $\Delta_{I-1} \times \Delta_{I-1} \times \cdots \times \Delta_{I-1} = \Delta_{I-1}^K$. By excluding loci that are monomorphic across the set of K populations, we exclude the I points $\mathbf{e}_1^K, \mathbf{e}_2^K, \ldots, \mathbf{e}_I^K$ in Δ_{I-1}^K in which all K allele frequency vectors $\mathbf{p}_k$ lie at the same vertex of the simplex.

The mean frequency of allele i across the set of populations is

$$\bar{p}_i = \frac{1}{K}\sum_{k=1}^{K} p_{ki}. \tag{1.14}$$

The homozygosity of the locus in population k is

$$J_k = \sum_{i=1}^{I} p_{ki}^2. \tag{1.15}$$

Considering the set of K populations simultaneously, homozygosity can be computed in two ways. The mean homozygosity across all populations is denoted by J_S,

$$J_S = \frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{I} p_{ki}^2. \tag{1.16}$$

The homozygosity of the population formed by pooling all K populations together into a single group is

$$J_T = \sum_{i=1}^{I} \bar{p}_i^2 = \sum_{i=1}^{I}\left(\frac{1}{K}\sum_{k=1}^{K} p_{ki}\right)^2. \tag{1.17}$$

Both J_S and J_T are functions of the set of allele frequency vectors $(\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_K)$, where $\mathbf{p}_k = (p_{k1}, p_{k2}, \ldots, p_{kI})$ is the allele frequency vector for population k. As in our descriptions of statistics of genetic diversity and genetic homogeneity, we drop the implied argument $(\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_K)$.

Suppose that a locus is polymorphic in the set of K populations. By definition of a polymorphic locus, $\bar{p}_i$ must be positive for at least two values of i, say $i = 1$ and $i = 2$. Neither $\bar{p}_1$ nor $\bar{p}_2$ can equal 1, as $\bar{p}_i = 1$ for some i would imply that $\bar{p}_j = 0$ for all $j \neq i$. Considering all polymorphic sets of allele frequencies in $\Delta_{I-1}^K \setminus \{\mathbf{e}_1^K, \mathbf{e}_2^K, \ldots, \mathbf{e}_I^K\}$, J_S lies in $(0, 1]$, and J_T lies in $(0, 1)$ (Exercise 1.4).

Perhaps the most commonly used statistic of genetic differentiation is F_{ST}, which, following [116, 118], can be defined as

$$F_{ST} = \frac{J_S - J_T}{1 - J_T}. \tag{1.18}$$

We can define the mean heterozygosity across the K populations and the heterozygosity of the full pooled set of populations by

$$H_S = 1 - J_S \tag{1.19}$$

$$H_T = 1 - J_T. \tag{1.20}$$

Writing F_{ST} in terms of heterozygosities rather than homozygosities, we have

$$F_{ST} = \frac{H_T - H_S}{H_T}. \tag{1.21}$$

Like J_S, J_T, H_S, and H_T, F_{ST} has implied argument $(\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_K)$. That F_{ST} is a sensible measure of differentiation follows from a result that $J_S \geqslant J_T$, which in turn follows from the Cauchy-Schwarz inequality.

Theorem 1.4: For all $(\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_K)$ in Δ_{I-1}^K, $J_S \geqslant J_T$, with equality if and only if $\mathbf{p}_1 = \mathbf{p}_2 = \ldots = \mathbf{p}_K$.

Proof: For a given i, consider the sequences of length K, $(p_{1i}, p_{2i}, \ldots, p_{Ki})$. By the corollary to the Cauchy-Schwarz inequality (Corollary 1.2),

$$K \sum_{k=1}^{K} p_{ki}^2 \geqslant \left(\sum_{k=1}^{K} p_{ki} \right)^2, \tag{1.22}$$

with equality if and only if $p_{1i} = p_{2i} = \ldots = p_{Ki}$.

Equation 1.22 applies for each i. Summing both sides of the inequality across values of i and dividing both sides by K^2, we have

$$\frac{1}{K} \sum_{i=1}^{I} \sum_{k=1}^{K} p_{ki}^2 \geqslant \sum_{i=1}^{I} \left(\frac{1}{K} \sum_{k=1}^{K} p_{ki} \right)^2. \tag{1.23}$$

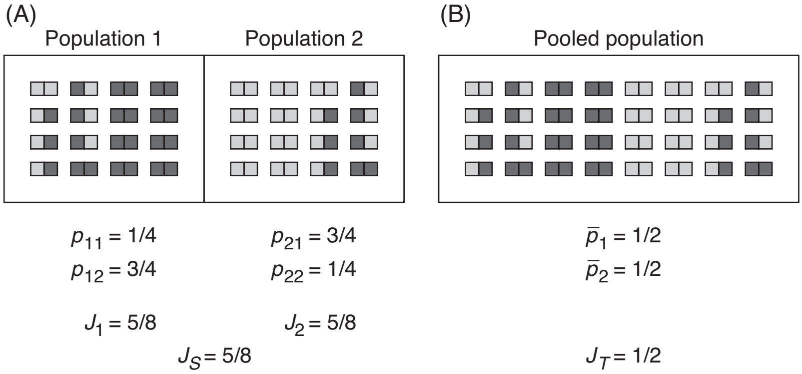

FIGURE 1.9. The quantities J_S and J_T for a pair of populations. Each pair of colored boxes depicts a diploid genotype for a locus with $I = 2$ alleles, with lighter boxes representing allele 1 and darker boxes representing allele 2. (A) Under Hardy-Weinberg proportions, each of the two populations has homozygosity $\frac{5}{8}$, so the mean homozygosity across the populations is $J_S = \frac{5}{8}$. (B) The pooled population obtained by averaging the allele frequencies of populations 1 and 2 has homozygosity $J_T = \frac{1}{2}$. The value of F_{ST} for the pair of populations is $(\frac{5}{8} - \frac{1}{2})/(1 - \frac{1}{2}) = \frac{1}{4}$.

Equality holds if and only if for each i, we have $p_{1i} = p_{2i} = \ldots = p_{Ki}$—that is, $\mathbf{p}_1 = \mathbf{p}_2 = \ldots = \mathbf{p}_K$. $\square$

This result, that $J_S \geqslant J_T$, provides a mathematical formulation of the well-known Wahlund principle, after [173], illustrating how a set of K populations has mean homozygosity across populations (J_S) greater than or equal to the homozygosity of the population considered as a whole (J_T) [138]. Indeed, a Wahlund principle holds for every allele (Exercise 1.5).

Because $J_S \geqslant J_T$, considering the bounds on J_S and J_T, considering all points in $\Delta^K_{I-1} \setminus \{\mathbf{e}^K_1, \mathbf{e}^K_2, \ldots, \mathbf{e}^K_I\}$, F_{ST} is bounded in $[0, 1]$. F_{ST} has a value of 0 if and only if $J_S = J_T$, which in turn requires all K populations to have the same allele frequency vector, $\mathbf{p}_1 = \mathbf{p}_2 = \ldots = \mathbf{p}_K$. As the K allele frequency vectors diverge, J_S becomes increasingly different from J_T, the value obtained for J_S if all K vectors are identical. Thus, because F_{ST} lies in $[0, 1]$, equaling 0 if and only if all K populations have identical allele frequencies and having larger values as the K allele frequency vectors diverge, F_{ST} is a sensible measure of genetic differentiation among K populations. The components of F_{ST} can be viewed in Figure 1.9.

F_{ST} traces to the work of Sewall Wright [186] on "fixation indices," as $J_S = 1$ implies that in each population, some allele is fixed at frequency 1 (not necessarily the same allele in each population). Note that we are using F_{ST} as a statistic computed from allele frequencies, a function that can be calculated from vectors of nonnegative values that sum to 1, with no consideration of a model that describes the evolution of populations. Another tradition in population genetics considers F_{ST} as a parameter of population-genetic models [73, 179, 180]. We will return to the topic of model-independence of allele-frequency statistics in Section 7.3.3.

1.6.2 Two Populations

In the case of $K = 2$ populations, genetic differentiation statistics are often viewed as measures of *genetic distance*, as they measure a "distance" between a pair of allele frequency vectors. The term "genetic distance" persists despite the fact that commonly used measures of the level of difference between pairs of allele frequency distributions do not necessarily satisfy all the mathematical criteria required for distance functions d. Such functions, on a set of points $\mathbb{M}$, must for each $\mathbf{p}, \mathbf{q}, \mathbf{r} \in \mathbb{M}$ satisfy

$$d(\mathbf{p}, \mathbf{q}) \geqslant 0, \text{ with equality if and only if } \mathbf{p} = \mathbf{q}, \tag{1.24}$$

$$d(\mathbf{p}, \mathbf{q}) = d(\mathbf{q}, \mathbf{p}), \tag{1.25}$$

$$d(\mathbf{p}, \mathbf{q}) \leqslant d(\mathbf{p}, \mathbf{r}) + d(\mathbf{r}, \mathbf{q}). \tag{1.26}$$

We use the term genetic distance as it is commonly used in the literature, without requiring that a genetic distance function satisfy all of these criteria. Typical genetic distance functions satisfy the first two criteria, but they do not necessarily satisfy the third criterion, the *triangle inequality* (Exercise 1.6) [10, 83].

For $K = 2$ populations, it is convenient to simplify the general expression for F_{ST} in eq. 1.18. The allele frequencies are p_{1i} for population 1 and p_{2i} for population 2. We write $\sigma_i = p_{1i} + p_{2i} = 2\bar{p}_i$ for the sum of the allele frequencies across the two populations, and $\delta_i = |p_{1i} - p_{2i}|$ for the absolute value of the difference. Each σ_i lies in $[0, 2]$, and each δ_i in $[0, 1]$.

Then

$$J_S = \frac{1}{2} \sum_{k=1}^{2} \sum_{i=1}^{I} p_{ki}^2 = \frac{1}{2} \sum_{i=1}^{I} (p_{1i}^2 + p_{2i}^2), \tag{1.27}$$

$$J_T = \sum_{i=1}^{I} \bar{p}_i^2 = \frac{1}{4} \sum_{i=1}^{I} \sigma_i^2. \tag{1.28}$$

We then have [24]

$$F_{ST} = \frac{\sum_{i=1}^{I} \delta_i^2}{4 - \sum_{i=1}^{I} \sigma_i^2}, \tag{1.29}$$

and in the $I = 2$ case, simply [140, 179]

$$F_{ST} = \frac{\delta_1^2}{\sigma_1(2 - \sigma_1)}. \tag{1.30}$$

A quantity D_{12}, which can be viewed as the homozygosity for a pair of sampled alleles taken from two different populations, appears in many formulas for genetic similarity and distance with two populations:

$$D_{12} = \sum_{i=1}^{I} p_{1i} p_{2i}. \tag{1.31}$$

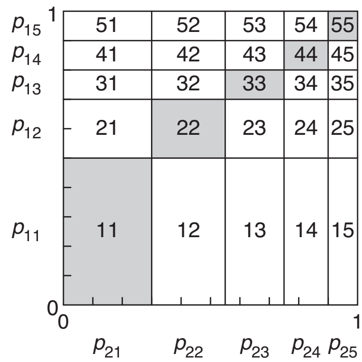

FIGURE 1.10. The allele-frequency dot product for a locus in two populations. Separately on the x- and y-axes, the unit interval is partitioned into components representing the frequencies of the different alleles in two populations. The allele-frequency dot product, represented by shaded areas, is $D_{12} = (0.5)(0.3) + (0.2)(0.25) + (0.1)(0.2) + (0.1)(0.15) + (0.1)(0.1) = 0.245$. Note that $J_1 = (0.5)^2 + (0.2)^2 + (0.1)^2 + (0.1)^2 + (0.1)^2 = 0.32$, and $J_2 = (0.3)^2 + (0.25)^2 + (0.2)^2 + (0.15)^2 + (0.1)^2 = 0.225$, so that $0.245 = D_{12} \leqslant \sqrt{J_1 J_2} \approx 0.268$.

This quantity, which we call the *allele-frequency dot product*, represents the probability that a pair of sampled alleles, one drawn from population 1 and the other drawn from population 2, have the same allelic type. The product $p_{1i}p_{2i}$ is the probability that both have type i, and D_{12} sums this probability across all allelic types (Figure 1.10). D_{12} is often denoted by J_{12}, or J_{XY} if populations are labeled by letters rather than numbers.

In terms of D_{12} and the homozygosities J_1 and J_2 of populations 1 and 2, $\sum_{i=1}^{I} \delta_i^2$ can be written as $J_1 + J_2 - 2D_{12}$, and $\sum_{i=1}^{I} \sigma_i^2$ can be written $J_1 + J_2 + 2D_{12}$. It then follows from eq. 1.29 that

$$F_{ST} = \frac{J_1 + J_2 - 2D_{12}}{4 - J_1 - J_2 + 2D_{12}}. \tag{1.32}$$

A second formula for F_{ST} that incorporates D_{12} can also be derived using eq. 1.18 together with the $K = 2$ cases of eqs. 1.16 and 1.17:

$$F_{ST} = \frac{J_T - D_{12}}{1 - J_T}. \tag{1.33}$$

The Cauchy-Schwarz inequality provides a simple relationship between D_{12} and J_1 and J_2.

Theorem 1.5: For all $(\mathbf{p}_1, \mathbf{p}_2)$ in Δ_{I-1}^2, $D_{12} \leqslant \sqrt{J_1 J_2}$, with equality if and only if $\mathbf{p}_1 = \mathbf{p}_2$.

Proof: The theorem is simply a restatement of the Cauchy-Schwarz inequality with $\mathbf{p}_1$ and $\mathbf{p}_2$ as the two sequences. The equality condition in the Cauchy-Schwarz inequality indicates that $D_{12} = \sqrt{J_1 J_2}$ if and only if $\mathbf{p}_1 = \lambda \mathbf{p}_2$. Because the entries of $\mathbf{p}_1$ and $\mathbf{p}_2$ each sum to 1 by eq. 1.1, $\mathbf{p}_1 = \lambda \mathbf{p}_2$ requires $\lambda = 1$. □

Table 1.1. Statistics introduced in this chapter.

Symbol	Concept	Equation
M	Frequency of the most frequent allele	1.3
J	Homozygosity	1.6
H	Heterozygosity	1.8
$J^{(\alpha)}$	α-homozygosity	1.9
J_A	Homozygosity with the first two alleles grouped	1.10
J_B	Homozygosity (equal to J)	1.11
J_C	Homozygosity excluding the most frequent allele	1.12
Z	Fraction of homozygosity not due to the most frequent allele	1.13
J_S	Mean homozygosity across populations	1.16
J_T	Homozygosity of a total population formed by pooling its subpopulations	1.17
F_{ST}	"Fixation index" measure of genetic differentiation	1.18
H_S	Mean heterozygosity across populations	1.19
H_T	Heterozygosity of the total population	1.20
D_{12}	Allele-frequency dot product	1.31
G	Nei's identity	1.34

This inequality in Theorem 1.5 motivates the definition of a measure of genetic similarity calculated from eqs. 1.31 and 1.15,

$$G = \frac{D_{12}}{\sqrt{J_1 J_2}}, \tag{1.34}$$

known as Nei's identity [115, 118]. By Theorem 1.5, for all pairs of allele frequency distributions representing polymorphic loci, Nei's identity lies in [0, 1]. If $\mathbf{p}_1 = \mathbf{p}_2$, then $G = 1$, as $D_{12} = J_1 = J_2$. A genetic distance $1 - G$ can be defined from Nei's identity; Nei's "standard genetic distance" (Exercise 1.7) is defined as $-\ln G$ [115, 118], and it lies in $[0, \infty]$. Nei's identity, denoted by G here, is often labeled I or I_{Nei}.

1.7 Do Statistics "Depend" on Allele Frequencies?

Now that we have introduced many of the main concepts (Table 1.1) and mathematical techniques for our analysis, we can pose the main questions answered by the book: in what ways do bounds on population-genetic statistics mathematically depend on the properties of allele frequencies, and how do these dependencies contribute to data analysis?

These types of questions have appeared in the literature: do population-genetic statistics "depend" on allele frequencies? Framed in this way, this question [61, 65] is not fully specified [98]. In what sense do we mean "depend"? Each population-genetic statistic—J, H, F_{ST}, and so on—is a function computed from allele frequencies and hence depends on the allele frequencies in a trivial sense.

The question can be reformulated in a more precise manner: how do properties of the allele frequencies constrain the values of a population-genetic statistic computed from those allele frequencies? Are the upper and lower bounds on a population-genetic statistic always the same, given the value of another statistic? In other words, does knowledge of the value of one statistic limit the possible values of another statistic?

We will see that if information is available about properties of an allele frequency distribution, then the values of those allele frequencies substantially constrain the values of population-genetic statistics. The upper and lower bounds of a population-genetic statistic are not in general identical across all values for another statistic: the value of one statistic limits the possible values of another statistic—often substantially. We establish these points by considering the relationships of a number of statistics to features of allele frequency vectors, as well as to each other.

We will also see that the constraints on population-genetic statistics as functions of other population-genetic statistics are important for the appropriate use of the various statistics. The dependencies among statistics generate correlations in the values of statistics computed from population-genetic data. In uncovering the mathematical bounds on population-genetic statistics, we illustrate their consequences for data analysis.

Each chapter focuses on a pairwise relationship between two quantities, loosely following the elements of a unifying pipeline (Section 7.2). Chapter 2 considers homozygosity and the frequency of the most frequent allele, drawing on [133, 139]. The following chapters consider four directions for extending results from Chapter 2. Chapter 3 examines a pair of statistics derived from homozygosity, drawing on [50]. Chapter 4 then studies homozygosity in relation to the frequencies of alleles subsequent to the most frequent allele. Chapter 5 discusses the extension of homozygosity to α-homozygosity, relying on the majorization method of [11]. Finally, Chapter 6 examines homozygosity and the frequency of the most frequent allele in finite samples, making use of ideas from [143] and results from Chapter 5. Chapter 7 summarizes the results, the techniques used to obtain them, and implications for data analysis.

Note that working with homozygosity is mathematically slightly simpler than working with heterozygosity, as we can examine sums of squares of allele frequencies without the additional step of subtracting them from 1. To view results in terms of diversity rather than homogeneity, homozygosity can be viewed as a diversity measure, where diversity increases with decreasing homozygosity, or results concerning homozygosity can be reframed in terms of heterozygosity by replacing J with $1 - H$.

1.8 Exercises

1.1 Suppose a locus has I distinct alleles. Prove that $J \geqslant \frac{1}{I}$, with equality if and only if $p_1 = p_2 = \dots p_I$.

1.2 Considering all possible allele frequency distributions in Δ_{I-1}, prove that $J_A \geqslant J_B > J_C$.

1.3 Explain why J_A and J_B are necessarily positive but J_C can equal 0.

1.4 Explain why J_S lies in $(0, 1]$ but J_T lies in $(0, 1)$.

1.5 In the proof of the Wahlund principle, interpret the meaning of the inequality eq. 1.22 in relation to the probability that allele i is homozygous. Use this inequality to introduce a stronger principle that has the Wahlund principle as a special case.

1.6 Provide a counterexample of three allele frequency vectors $\mathbf{p}, \mathbf{q}, \mathbf{r}$ that demonstrate that F_{ST} does not satisfy the triangle inequality.

1.7 Does Nei's standard genetic distance satisfy the triangle inequality on Δ_{I-1} for $I \geqslant 2$? Provide a proof or counterexample.

AUTHOR INDEX

SUBJECT INDEX