PREFACE

In the play *Arcadia* by Tom Stoppard, the mathematical population biologist Valentine Coverly wrestles with an issue that pervades the mathematical study of populations: the nature of the relationship of the mathematics to biological phenomena in the real world. Using historical data on the grouse population at an English country house, the character seeks to find a mathematical rule governing population dynamics. The play refers to the potential generality of the equation that is sought, say, for populations of goldfish in a pond, among other examples (p. 45). But Valentine stumbles on the meaning of the task: even if a suitable rule can be found, what exactly is the mathematics describing? The idealized population in the mathematical model, or the population in the real world?

In mathematical population genetics, the answers to the questions raised by Valentine's quest can usually be discerned. The population geneticist works out the consequences of mathematical assumptions within the context of a mathematical model. The assumptions of the model are inspired by biological phenomena, but the mathematics that follows describes the mathematical entities that the model creates, not actual biological populations. It is hoped that real-world populations are approximated by corresponding entities in the mathematics, and biological insight is derived by juxtaposing the math with the biology.

In empirical population genetics, which proceeds from computations performed on population-genetic data rather than inside a mathematical model, the role of mathematics is not as clear. The empirical study of genetic variation in groups of organisms—biological populations—begins with summary statistics describing that variation. These statistics, often computed from frequencies of alleles in the populations, are then interpreted to make claims about the biological populations. The biological interpretations rely on the large body of mathematical theory about the behavior of the summary statistics that has been developed throughout the rich tradition of mathematical evolutionary modeling, often alongside simulations of population models and a long record of computation of the statistics. A researcher might claim from the statistics that a population is currently experiencing some specific biological phenomenon, such as growth in size, an exchange of migrants, or inbreeding. The inferential step is not itself mathematical, but, as we will see, the mathematics remains present, if hidden.

Mathematics of population-genetic summary statistics. Population-genetic summary statistics capture basic aspects of the features of genetic variation, and they serve as the cornerstone of data analysis for such fields as conservation genetics, human population genetics, and molecular ecology. Even as the broader field of population genetics has devised new forms of statistical inference based on rich techniques from modern

computational statistics, summary statistics remain at the center of population-genetic data analysis. Summary statistics provide initial observations and first-order inferences, and state-of-the-art statistical methods are then used only alongside them. The summary statistics often suggest choices for deeper analyses and aid in the interpretation of their results. The role of summary statistics in population genetics can be compared to the basic skills of learning to play a sport or a musical instrument: a mastery of fundamentals is important for advancement with more complex techniques.

Often, however, even the most foundational statistics in population genetics generate results that are counterintuitive based on an understanding of how they have been designed. For example, in testing the fit of population-genetic models to genetic data, two statistical tests relying on seemingly different quantities—the frequency of the most frequent allele, which focuses on the frequency of only a single highly frequent allele, and homozygosity, which considers *all* of the allele frequencies—might be expected to uncover deviations from the models under different conditions. Yet statistics using the two quantities can instead produce nearly identical behavior. In other examples, a population-genetic statistic—perhaps the genetic differentiation statistic F_{ST}, perhaps a ratio of homozygosities—has been designed to range from 0 to 1, so that 0 and 1 represent the most extreme theoretical possibilities. Yet the statistic is rarely observed to have values near one or both ends of the unit interval, even in scenarios regarded as biologically extreme.

This book argues that despite the central role that summary statistics have long played in population-genetic data analysis, one key element in the mastery of these fundamentals of population-genetic data analysis has been underappreciated. In particular, observations such as the limited range of statistics that are supposed to span the full unit interval and the near-identical behavior of seemingly distinct statistics arise not from biological phenomena of interest to the population geneticist, but, rather, from often-overlooked mathematical properties of the statistics themselves. As we will see, these examples illustrate that the mathematical features of statistics computed from allele frequencies—sets of nonnegative numbers that sum to 1—can interfere with proper biological interpretations. An understanding of the mathematical properties of population-genetic statistics can then help in revealing new connections among different statistics, designing new approaches to data analysis, clarifying the biological inferences that are justifiable from patterns of genetic variation—and enhancing the understanding of the applications and limitations of the field's most fundamental tools.

Hence, the book seeks to systematically uncover the influences of the mathematics of allele frequencies on population-genetic statistics, and to illustrate how an understanding of the mathematical properties of population-genetic statistics enhances the interpretation of genetic variation in populations. The focus is on a single class of statistics, those derived from the concept of *homozygosity*. The key issues in the mathematics of population-genetic statistics all emerge with this class of statistics based on a simple quadratic function of allele frequencies, a "quadratic form." These include: the mathematical dependencies of pairs of statistics, the tools for uncovering those dependencies, and the potential approaches to addressing them—all of which can be studied with homozygosity statistics. After detailed analysis of homozygosity statistics, the reader will not only be better equipped for interpreting mathematical features of statistics based

on homozygosity, but will also have a framework for applying the book's techniques to explore relationships among other allele-frequency-based statistics.

The book's approach. The approach of the book is mathematical—employing derivations of theoretical bounds for certain population-genetic quantities in relation to the values of other such quantities—with a view toward the behavior of statistics in the empirical analysis of population-genetic data. After an introduction to concepts and questions in Chapter 1, Chapters 2–6 each center on the relationships among two or more statistics. Theorems are derived to provide formal mathematical relationships among the statistics, often in multiple settings with different constraints. The mathematical results are illustrated in population-genetic data, and the reader is invited to derive ancillary results in the exercises. Chapter 7 collects many of the results and techniques, summarizes additional results in the spirit of the book, and expands the argument for the merits of the overall approach.

A work at the intersection of mathematics, statistics, and empirical population genetics necessarily draws readers from heterogeneous backgrounds. Although readers from one specific area may find stylistic elements common in other areas to be peculiar, the text seeks to appeal to diverse readers. The work is situated in the tradition of mathematical population biology, in which (i) scenarios motivated by questions in biology are introduced, (ii) mathematical problems are then presented and solved, and (iii) the mathematical solutions are then interpreted in relation to motivating scenarios. For readers approaching the book from this tradition or from related areas of computational biology, the book can provide a map to further mathematical research on population-genetic statistics and theory for population-genetic data analysis.

For mathematical readers with little biology background, such as mathematics students early in their engagement with mathematical biology, the book offers a detailed study of a biological area in which elementary mathematics—primarily calculus and inequalities—has much to contribute. Though such readers will certainly benefit from knowledge of population genetics, little background in population genetics is assumed, and the book offers a portal to an intersection of math and biology using only minimal biological terminology. Mathematical readers might enjoy a focus on the interiors of chapters and on the exercises, the more difficult of which are starred.

For readers situated in empirical population genetics and related areas of molecular ecology and conservation genetics, the book offers a fresh perspective on familiar statistics. The mathematics is generally no more advanced than calculus, providing a basis for using the book without extensive mathematical background. However, such readers might only wish to analyze the more technical material in the interior of a chapter after a close look at its beginning and end. Chapter 7 compiles the main technical results in a systematic discussion of the use of population-genetic statistics; a biological reader might focus on the last chapter, tracing back through the book to sections that substantiate its points.

Finally, although the emphasis of the work is on allele-frequency data in population genetics, many of the same mathematical functions used for analyzing allele frequencies appear in other fields, including ecology, economics, and health services research. For example, readers familiar with the measurement of biodiversity will find echoes of discussions of ecological diversity statistics, and, often, mathematical results that characterize

features of these statistics. For statistical readers interested in indices that summarize categorical data in ecology and other areas, population-genetic language is kept to a minimum, with mathematical results described in terms of vectors that sum to 1—so that results can be imported mutatis mutandis to other fields.

At the book's end, the reader will appreciate how the values computed for summary statistics are sometimes, as Valentine Coverly says about the values produced by his population dynamics equation, "not about the behaviour of fish" but, rather, "about the behaviour of numbers." Through a mastery of a variety of bounds on population-genetic statistics, the reader will possess an enriched framework for making biological inferences from summary statistics nevertheless.

Quadratic forms most beautiful. Charles Darwin's famous conclusion to the *Origin of Species* marvels at the diversity of life—the "endless forms most beautiful and most wonderful" that "have been, and are being, evolved." Owing to the discreteness of categorical data and the consequent appearance of floor and ceiling functions in mathematical bounds on population-genetic statistics, this book finds that from so simple a beginning—a quadratic function of allele frequencies—graphs of those statistics turn out to possess unexpectedly intricate shapes. It is hoped that in addition to finding the bounds and resulting graphical regions scientifically valuable, readers will also delight in their beautiful forms.