Chapter 1

Introduction

1.1 Why this textbook?

Welcome to **Biostatistics: Data and Models**! This textbook provides a survey of statistical methods commonly used in the life sciences, an introduction to statistical theory, and significant exposure to the statistical software package SAS 9.4 (©2020 SAS Institute Inc.). The textbook is designed for graduate students and upper division undergraduates in the life sciences, and assumes some familiarity with mathematical notation, functions, and algebra. A review of these topics is also presented early in the textbook. Knowledge of calculus is helpful but not essential. No previous courses in statistics are needed.

There are many useful introductory statistical textbooks (e.g., Sokal & Rohlf 1995, Steel et al. 1997, Schork & Remington 2000), so what is different about this one? One is the close integration of the text with SAS programs and output. Many texts do not discuss a particular software package, provide only abbreviated examples, or present them under separate cover. However, these packages play an essential role in modern statistical analyses, and fluency in a statistical language is a basic tool for the practicing scientist. I selected SAS as the statistical package for this textbook because of its popularity, extensive documentation, and strong support of mixed models, a common statistical procedure. An alternative is the free software package called R (R Core Team 2021). For those interested in learning this software, R programs similar in function to the SAS code can be downloaded at the website for this text.

Another difference in this textbook is the integration of statistical procedures and theory. Most introductory textbooks present the statistical procedures and a mechanistic explanation of how they work, without discussing the underlying theory. The theory is typically presented in advanced courses to a more mathematically inclined audience. However, I feel that some knowledge of the theory is essential for students in the life sciences, and so some theoretical concepts are included in this text. For example, likelihood is used throughout the text to explain how parameters are estimated and statistical tests derived. Besides many basic statistical procedures, likelihood theory also plays a role in model building and selection using information criteria, as well as Bayesian statistics, an expanding field of statistical analysis.

As part of this integration of theory, statistical models are presented throughout the text. What is a statistical model? Suppose we are interested in fitting a line through some data points, which are in the form of (Y, X) pairs. A standard statistical model for fitting a line through such data is the linear regression model:

$$Y = \alpha + \beta X + \epsilon, \tag{1.1}$$

where α is the intercept of the line, β is the slope, and ϵ represents random variation of the data around the line. **There is always some random variation around the line, especially with biological data.** If there were no random variation, a statistical approach would not be needed – one could simply draw a line through the data.

Fig. 1.1 shows an example of this model, fitted to data on the number of reptile species on islands of varying size in the West Indies (Wright 1981). We will examine how the parameters of such models (α and β) can be estimated using likelihood theory, and how to test whether there is indeed a relationship between Y and X (as it appears in Fig. 1.1). It is also possible to make predictions from statistical models. For example, we could use this model to potentially predict the number of reptile species expected on other islands, ones not included in this data set.

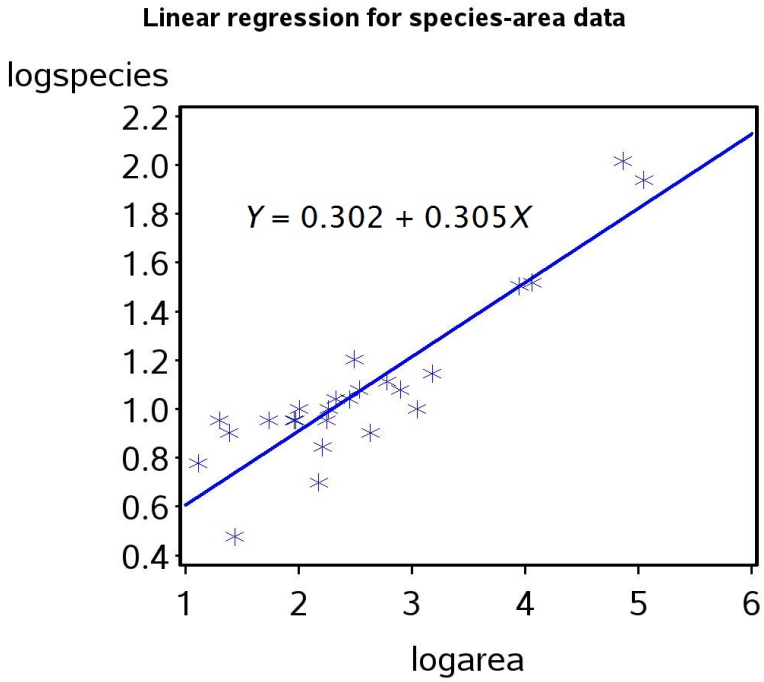

Figure 1.1: Number of reptile species vs. island area in the West Indies (Wright 1981). The number of species and island area were log-transformed before analysis. The fitted line and model equation are also shown.

1.2 Types of data

The first step faced by the statistical analyst is determining the form of the data. There are four types of data frequently encountered by scientists and statisticians: continuous, discrete, rank, and categorical data. **Continuous data** are quantities like the length and weight of an organism, concentrations of chemicals in the environment, or the growth rate of a population. The distinguishing feature of continuous data is that the observations can be described using real numbers. For example, the length of an organism might be 4.53 cm, while its weight 1.23 g. In contrast, **discrete data** always take integer values. They can be counts of organisms in a location, the number of vertebra in the spine, or quantities like the number of disease cases in a month. Typically, discrete data are non-negative integers, i.e., $0, 1, 2, 3, 4$ and so forth. The number of species in Fig. 1.1 could be treated as either continuous or discrete - although they are integer values, they are large enough to take many potential values and approximated as continuous data.

Rank or ordinal data are observations that indicate the relative ordering of the data. For example, suppose an entomologist wants to rapidly assess the level of damage caused by caterpillars to their host plants. It may be easy to quickly assess whether the plants have no damage (a rank of 1), or light (2), medium (3), and heavy damage (4), but finer gradations would be difficult. Rank data also play an important role in a set of procedures called nonparametric statistics, because these procedures often convert continuous or discrete data to rank data. **Categorical data** are observations that fall into separate categories. For example, we might classify specimens of an animal as male, female, or juvenile. No numbers are associated with these categories, although we would likely be interested in how many animals occur in each category, i.e., their frequencies.

1.3 Data and models

Once the data are classified into one of the above types, this determines to a large extent the statistical analysis. For example, suppose the data are (Y, X) pairs as in Fig. 1.1. A linear regression model like Eq. 1.1 would seem appropriate, because the data lie near a straight line. How could we model the random variation around the line? One common choice is

to assume that ϵ has a normal or bell-shaped distribution, which we later examine in detail. Once a statistical model is chosen, this largely determines the analysis including how model parameters (α, β, and parameters for ϵ) are estimated and statistical tests conducted, often using likelihood theory. Another important task in statistics is model building, in which a number of different models are fitted to the data and the best-fitting one selected (there are various criteria for determining which is best). Fig. 1.2 shows this general process.

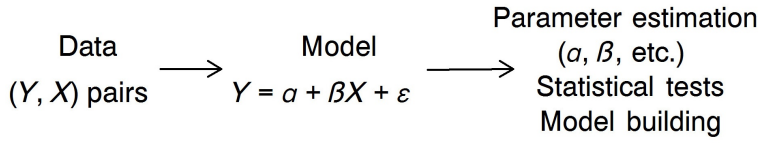

Figure 1.2: Sequence of analysis for many statistical problems.

1.4 Sequence of topics

The next chapter in this text is a brief review of the mathematics useful in statistics, and an introduction to SAS programming (Chapter 2). We then introduce descriptive statistics, which are quantities like the mean or average, designed to summarize the properties of a data set (Chapter 3). The next topic is probability theory, which provides an explanation for many natural processes that apparently have random components, and provides a foundation for statistics (Chapter 4). We then turn to probability distributions for both discrete and continuous data, which are essentially models for random processes (Chapter 5 and 6), and how means and other quantities are defined for these distribution (Chapter 7). We then examine how parameters for these distributions are estimated using likelihood, along with a measure of the reliability of these estimates (Chapter 8, 9), and how hypotheses concerning the parameters are tested (Chapter 10).

Several chapters are devoted to analysis of variance, or ANOVA, used to compare the means of different groups (Chapter 11-15). These groups are often generated by different experimental treatments, and ANOVA and related

techniques provide a way of examining whether the treatments produces differences among these groups. Nonparametric alternatives to ANOVA are also considered (Chapter 16). We then examine linear regression and correlation, which are alternate methods of examining the relationship between two variables (Chapter 17, Chapter 18). These methods are designed for continuous variables, but can be adapted to discrete ones. Chapter 19 presents more complicated designs including three-way and nested ANOVA, and analysis of covariance (ANCOVA). In Chapter 20, we examine several techniques useful for analyzing categorical data. Chapter 21 provides an introduction to multiple regression, which examines how one continuous variable is affected by several other variables. Several large data sets used as examples are listed in Chapter 22, while Chapter 23 contains statistical tables used throughout the text.

1.5 References

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. The R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Schork, M. A. & Remington, R. D. (2000) *Statistics with Applications to the Biological and Health Sciences, Third Edition*. Prentice Hall, Upper Saddle River, New Jersey, NJ.

Sokal, R. R. & Rohlf, F. J. (1995) *Biometry, Third Edition*. W. H. Freeman and Company, New York, NY.

Steel, R. G. D., Torrie, J. H. & Dickey, D. A. (1997) *Principles and Procedures of Statistics: A Biometrical Approach, Third Edition*. McGraw-Hill, Boston, MA.

Wright, S. J. (1981) Intra-archipelago vertebrate distributions: the slope of the species-area relation. *American Naturalist* 118: 726-748.