CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION AND OVERVIEW

Statistics is viewed by many as a branch of mathematics and since math is widely regarded as a difficult subject, people often assume automatically that statistics must be just as difficult. In truth, there is much not to "fear" in statistics, which in some cases involves nothing more complex than elementary arithmetic operations. As proof that statistics is not as "unappetizing" as it might seem, it is one of the few disciplines that find useful applications across a wide variety of professions, including the medical and biomedical, social sciences and psychology, economics, environmental and engineering professions, and so on. The environmental profession is almost completely driven by data, but surprisingly does not seem to have embraced statistics as much as some of the other professions just mentioned. Environmental engineers, geologists, scientists, or other professionals engaged in miscellaneous practice, research, or related activities have to work routinely with data from soil and sediment, surface and groundwater, ambient air, and other environmental media, for a variety of reasons or purposes. Data are collected to investigate or monitor an environmental concern, analyzed, and interpreted to gain insights into the situation, and the findings used to support decisions on response actions or predict future outcomes.

There appears to be a number of possible reasons why the use of statistics is not as widespread among environmental professionals as could have been expected. One reason is that it is largely a regulated profession, and the (government) regulators often "usurp" statistical responsibility for the data analysis. That is, environmental statutes and regulations regarding pollutants typically incorporate numerical standards and requirements that the regulated community is expected to abide by, and those numerical standards and criteria

Statistical Applications for Environmental Analysis and Risk Assessment, First Edition. Joseph Ofungwu.

typically incorporate some form of statistical analysis and estimation procedures. Therefore, technically, the environmental professional only has to follow the regulators' guidelines, and not be unduly burdened with performing miscellaneous statistical analyses on his or her data. The problem with this minimalistic approach obviously is that the regulatory standards are necessarily generic (i.e., "one size fits all"), since the regulators cannot possibly have foreknowledge of every conceivable site or circumstance. Usually, the regulations allow for alternate site-specific standards that more realistically reflect the particular site conditions and characteristics if supported by the appropriate statistical analysis, but in practice, many practitioners decline to develop such alternate standards when feasible (see the next paragraph for possible reasons), opting instead to use the generic standards reflexively. As a result, decisions on response actions are often reached that are overly conservative for some sites (e.g., waste of resources for unnecessary cleanup actions), or insufficiently protective for other sites (e.g., chemical exposure risk falsely determined to be within acceptable limits).

A second likely impediment to widespread use of statistics among environmental professionals is that the college curricula for the associated disciplines (environmental engineering, geology, environmental science, etc.) sometimes do not include or emphasize environmental statistics, while standard statistical books are often not sufficiently readable and sufficiently relatable for readers without a mathematics background to fully appreciate. A third related reason is the high cost of high-quality statistical software. Although software cannot substitute for functional familiarity with the basic statistical concepts, access to software does provide a powerful incentive to get more involved and cultivate the habit of subjecting the data to greater analysis. Nothing kills the appetite like trying to perform multiple nonlinear regression by hand! The more affordable software packages are often selective in what tasks they can or cannot perform, which can become quite frustrating. The high-quality commercial software systems that do it all can be prohibitively expensive and difficult to justify, especially by someone who is not even at the level yet to understand what to do with or expect from these high-cost software packages. Fortunately, the advent of the freely available R software system, with its extensive and ever growing range of capabilities, has taken the constraint of software access off the table.

1.2 THE AIM OF THE BOOK: GET INVOLVED!

Given the above background, the main purpose of this book is to present fundamental statistical principles and procedures, in the context that they are commonly encountered in environmental practice, in simple and unambiguous language. The powerful free software package, R, plays an important supporting role in this regard. As indicated above, manual computation of many statistical tests is impractical due to the level of effort required, and fully functional software is not always affordable. With the easy availability of R, all that is really needed is a good grasp of the basic statistical concepts, as described in this book, and R can help take care of the rest. No excuse to wait any longer! However, although almost unlimited in its functionalities and with an ever growing list of user-contributed packages and procedures, R is not completely without issues. For one, it does not have the point-and-click ease of use that is usually the norm with commercial software, but the good thing is that there are "portable" scripts as provided throughout this book that will readily compute the desired analyses or procedures. The reader only needs to substitute his or her own data for the data used in the worked examples in the book, and document the scripts for future reuse as necessary. Trying to become a certified R expert can indeed be daunting, but taking it in small bite-sized scripts and macros can be surprisingly easy, always keeping in mind that for each function or command that performs a particular task, there are other alternative

functions that can perform the same task and even better, with more functions becoming available as R continues to expand. Another potential hurdle is that there is no dedicated technical support service for R, although there are mailing lists and kind souls who usually respond to user requests for assistance out of the goodness of their hearts.

There are obviously other software packages besides R, including freely available as well as commercial packages, and since no software system is necessarily infallible, whenever in doubt, it is prudent to compute the same analysis using more than one software system, if available. For these reasons, alternative software packages besides R that are used for computing the numerous worked examples in this book include the freely available ProUCL, Visual Sample Plan (VSP), and DATAPLOT software packages, as well as the popular and affordable commercial *Minitab*® (version 16). ProUCL Version 4 has received criticism in the literature for deficiencies in some areas (see Chapter 2), but is still a very useful software package especially for the analysis of data containing nondetects (NDs). Note that Version 5 was recently released and may have addressed many of Version 4's shortcomings. Minitab is a modestly priced, conveniently user-friendly software product that would be suitable for the reader who does not wish to experiment with or depend solely on free software. The objective is not only to provide sufficient software options that the reader is unlikely to be hindered by lack of access to software, but also to heed the conventional wisdom of not putting all our (statistical) eggs in any one basket.

As indicated above, environmental practice is typically subject to regulatory oversight, and regulatory departments usually have access to professional statisticians. Therefore, it is crucial for the environmental professional who wishes to engage in statistical communication with regulatory officials to have full confidence in his or her grasp of at least the basic concepts. It is the intent in this book to provide the necessary level of detail and illustration to foster that confidence.

A related objective of the book is to draw attention to the critical but often overlooked role of statistics in human health and ecological risk assessment. The book is divided into four parts and the last part (Part IV) links the various statistical techniques and procedures of the preceding parts to the ultimate objective of assessing contaminant exposure risk. It is often the case that the analyst is performing a great many statistical tests and procedures, but not always keeping in focus the end purpose of all the analyses. Why are we comparing populations, the background and the site, or interested in the true mean concentration of a contaminant, or exploring contaminant trends and patterns? The exposure point concentration (EPC), which is the concentration of the contaminant or chemical that is assumed to contact the human or ecological receptor, is probably the most important exposure factor driving the risk assessment, and determination of the EPC is mainly a statistical task.

The bulk of environmental practice revolves around contaminant data collection and analysis, for the purpose of determining whether the contaminants pose unacceptable public health or ecological risks warranting corrective action, or if the exposure risks are minimal. Unfortunately, in many cases, risk assessors are toxicologists by profession and often focus more on the toxicity properties of the chemicals (unit risk factors, reference concentrations, etc.) while glossing over the statistical aspects of risk assessment. By introducing elements of risk assessment in Part IV, this book highlights the important role of statistics in protecting public and environmental health, while avoiding unnecessary environmental cleanup costs.

1.3 THE APPROACH AND STYLE: CLARITY, CLARITY, CLARITY

It is discouraging when the chapter material in a book is so full of puzzles and riddles (wittingly or unwittingly) that the reader is mentally exhausted by the end of each chapter

and has no confidence left to even read, much less tackle, the chapter exercises. I cannot imagine that the average reader would not prefer that the text material is explicit, so that there is no need for guesswork as to what the author has in mind, while the "hide and seek" is reserved for the end-of-chapter exercises. Accordingly, clarity of communication is the overarching stylistic goal in this book, even if the text is sometimes repetitive or awkward or the page count is a bit higher, as a result. The idea is to make it nearly impossible to misunderstand or have to guess the intent of anything being stated. Along the same lines, I have endeavored to enhance readability by minimizing the use of acronyms (or frequently spelling them out), and frequently providing cross-references to the locations of helpful sections and chapters for the convenience of the reader who is not reading the book in sequence from the beginning to the end (does anyone ever?). To the advanced reader who might prefer faster-paced text, all of this may sound boring; so, your kind indulgence is respectfully requested! I truly believe that if a book is intended for a general audience, it is incumbent on the author to continually keep in mind the nonexpert members of that audience (i.e., those who really need the book in the first place!), whose needs for a little more detail should always drive the writing style.

The book is divided into four parts. Part I describes basic statistical measures and concepts, including graphics. Part II describes statistical procedures for univariate data, while Part III focuses on regression and other multivariable data analyses. Regression is probably the most frequently used statistical analysis, with entire books written not just on the subject but on various aspects of it as well. Obviously, it is impossible to cover every aspect in every detail in any one book. Instead, this book focuses on the more commonly used regression methods, but provides the necessary context and basis for the interested reader to obtain additional information, if needed. Part IV outlines the role of statistics in environmental data collection and subsequent exposure risk assessment. Each chapter begins with an "Introduction and Overview" section that lays out the chapter material in one or two pages on average, to orient the reader and provide the necessary background. Subsequently, the chapter material is presented, liberally interspersed with worked examples using multiple software applications and spreadsheet computations where feasible, to fully illustrate the concepts. Some of the worked examples are rather lengthy, as they include appropriate commentary on various aspects of the results derived. Since the reader's data and results will most likely be different from those described in the examples, I believe it is important to sufficiently describe the process for the reader to be able to adapt and apply with confidence.

Finally, the pesky but very common problem of left-censored data (i.e., nondetects or NDs in the data sample) is always addressed. Environmental contaminant data samples frequently include NDs, where an ND simply implies that the exact contaminant concentration is unknown, but lies anywhere between zero and the laboratory detection limit or reporting limit of the contaminant in question (see Chapter 3 for detection and reporting limits). Without knowing the value of the concentration, how can it be included in the various statistical analyses? Even worse, NDs may deceptively seem like a benign sort of problem because excessively high contaminant concentrations would normally be of greater concern than concentrations that are so low as to be undetectable, but in reality, improper handling of NDs can skew the results of some statistical tests, among other potential consequences. The important work by Helsel and others (Helsel, 2012) provides numerous options for dealing with left-censored data other than the commonly used but undesirable practice of arbitrarily substituting zero or some other assumed value in place of the unknown NDs.