1

Introduction

1.1 The textbook's purpose

The book *Improving Natural Resource Management: Ecological and Political Models* (Haas, 2011), in pages 206–207, provides a list of areas that ecosystem managers need stronger training in. This textbook provides such instruction in probability, statistics, simulation, and resampling methods. It is designed to be used either as a classroom textbook or as the main text supporting a self-study regime (with web-based aids) by practicing ecosystem managers who are unable to attend an instructor-led course.

Topics salient to ecosystem management, but rarely covered in introductory statistics textbooks, are introduced. These include

1. basics of Bayesian networks and influence diagrams;
2. minimum simulated Hellinger distance (MSHD) parameter estimation;
3. resampling-based hypothesis tests and confidence intervals that can be applied to spatio-temporal data;
4. spatial and spatio-temporal statistics;
5. learning algorithms for influence diagrams; and
6. individual-based models (IBMs) of wildlife populations.

Influence diagrams are Bayesian networks that have decision and utility nodes. IBMs are also called *agent-based models*.

Probabilistic and statistical algorithms taught in the textbook are applied to wildlife management problems. Specifically, the reader will see how to construct

Introduction to Probability and Statistics for Ecosystem Managers: Simulation and Resampling, First Edition. Timothy C. Haas.

and run influence diagrams of the decision making of ecosystem stakeholders and IBMs of wildlife population dynamics.

IBMs are increasingly being used to model complex interactions between wildlife, humans, and the landscape. There are few introductory treatments of IBMs and fewer that integrate an introduction to IBMs within an introductory probability and statistics textbook. In addition to several R codes aimed at the introductory level, the **id** software system that accompanies this textbook supports the construction and use of wildlife population IBMs.

As in Haas (2011), the present textbook shows how to build models of the decision making of an ecosystem's legally sanctioned stakeholders and those groups of stakeholders operating outside the legal limits of the country or countries hosting an ecosystem. Such stakeholders are aggregated into roughly homogeneous social groups or simply, *groups*. Ignoring intragroup dynamics, the decision making of such a group is viewed herein as that of a single macro-individual. Such a model is realized as an influence diagram of a group's decision making as regards other groups and the ecosystem. Hereafter, the influence diagram model of a group's decision making will be referred to as a *group decision-making diagram*. For example, in Chapter 2, decision-making models are developed for two South African groups: rhino poachers and policymakers. These models are referred to as the *rhino poachers' decision-making diagram* and the *policymakers' decision-making diagram*, respectively.

Allowing such group decision-making diagrams to learn from experience was listed as a future research area in Chapter 13, pp. 204–205 of Haas (2011). Adding such a capability to the EMT of Haas (2011) was also suggested by a reviewer of that textbook (Filatova, 2011). Such algorithms, described herein, have been developed and added to the **id** software system (Haas, 2011).

1.1.1 The textbook's focus on ecosystem management

The following characteristics make this textbook particularly effective at helping ecosystem managers learn basic probability and statistics.

1. Ecosystem management, for example, the management of wildlife populations, control of invasive species, or the management of environmental pollution is increasingly based on large simulation-based models. However, most introductory textbooks contain little instruction on such models.

2. Simulation is an accessible way to explain probability and stochastic model behavior to beginners. Rosenshine *et al.* (2002) refer to *axiomatic* versus *simulation* approaches to teaching probability. These panelists see students building a more intuitive and experience-based understanding of probability through the use of simulation in their exercises. This approach to teaching probability is addressed in more depth in the last section of this chapter.

3. Statistical inference on models that do not have analytical likelihoods is common in complex ecosystem management modeling. In these cases, resampling-based statistical tests are often the only alternative available to the ecosystem manager. However, little instruction is available on performing such statistical inference that is accessible to beginners.
4. Bayesian networks provide a graphical way to teach introductory probability and statistics, but few introductory textbooks take this approach.
5. Ecosystem managers almost always face a spatio-temporal situation that is stochastic. Few introductory textbooks incorporate such topics. Usually, introductory GIS books consider spatial-only statistics. This textbook introduces spatio-temporal statistical models.
6. Self-study is often the only option for practicing ecosystem managers, but most textbooks are intended for classroom use. Coupled with its accompanying OITS, this textbook is tailored for the reader who finds himself or herself without the aid of an instructor.
7. The main goal of this textbook is to introduce probability and statistics to future and practicing ecosystem managers by providing enough explanatory content so that such a reader can fully grasp the nature of these two areas. Following these explanations, applications of probability and statistics are selected with no intention of being exhaustive but rather with an eye toward (i) supporting the author's first textbook, which is mentioned in the beginning of the chapter, (ii) showing the proper use of probabilistic modeling and computationally intensive statistical inference, and (iii) providing introductory instruction in modeling and inferential tools that are becoming increasingly important for the management of political–ecological systems.

1.1.2 Reader level, prerequisites, and typical reader jobs

This textbook assumes that the reader has a working knowledge of algebra and has taken a course in pre-calculus. No calculus, however, is needed. The vectors and matrices used in this textbook are explained in detail at their point of introduction (the beginning of Chapter 6).

The textbook is appropriate as the first course in ecosystem (wildlife and/or environmental) modeling in an undergraduate program. It is also appropriate for an online course or for self-study.

The book will be particularly helpful to a reader who is (i) a student in a natural resources management program; (ii) an analyst, project leader, or manager within a forestry, fisheries, national park, environmental protection agency, or other conservation-focused government agency; or (iii) an employee of a wildlife advocacy organization such as the African Wildlife Foundation or the World Wildlife Fund.

1.2 The textbook's pedagogical approach

1.2.1 General points

The pedagogical viewpoint of this textbook is that explanations should be brief, mathematical symbols should be emphasized, and many exercises should be computer based. The text's examples are taken from the field of ecosystem management.

One theory in education states that a concept is learned as a person is able to recognize it and use it in a variety of situations and contexts. To do this, it is essential to vary the words used to describe the concept and to give practice in the use of the concept in apparently disparate situations.

This text's associated software files are accessible through the author's website, `www4.uwm.edu/people/haas/introtext/` This textbook contains many exercises. Solutions to these exercises are available online at this URL.

An accompanying OITS written by the author and tuned to this textbook is available at `www4.uwm.edu/people/haas/introtext/oits/oits.html` This system can help a reader master the textbook's material – especially one who is intending to study the textbook without the aid of an instructor. Such tutoring systems are currently rare among introductory probability and statistics textbooks.

1.2.2 Use of this textbook for self-study

A criticism of many self-study systems is that they are boring. An ecosystem manager trying to learn a concept through a self-study system does not have the benefit of peer pressure and/or encouragement from his/her classmates to study. Further, a reader working alone does not have the stimulation of contact with an instructor as with a traditional, face-to-face course. Hence, the challenge to motivate thinking must come entirely from the textbook and accompanying online aids. If the ecosystem manager finds the textbook's exercises too easy, he or she will lose interest in learning. Hence, contrary to many self-study programs, the approach taken herein is to balance straightforward, skill-building exercises with several difficult and challenging problems for the ecosystem manager to solve. Although fully worked out solutions are available online, if a complete solution is read before attempting an exercise, the motivation to seriously think through to a solution may be killed.

The best exercises are those that are self-checking, that is, after a solution has been worked out, some mathematical or computational procedure can be applied to at least partially check the answer's correctness. For example, any probability answer must be in the unit interval. Such diagnostics accompany many of the exercises herein.

Many self-study systems use random generation of new datasets within exercises to supposedly keep the exercise from being predictable. This will be avoided here because emphasis will be on concept understanding and not on the mechanics of data analysis.

Learning takes place during a serious effort to solve an exercise. This is true of any quantitative subject regardless of the presentation medium. Hence, the completion of the exercises is viewed here as the ecosystem manager's most important learning activity.

Another criticism of self-study systems is that it is tedious to read through reams of text – either printed or online. This textbook avoids this pitfall by keeping explanations and presentations of new material to a minimum number of words needed to convey the idea. Brevity is a high priority in what follows.

Probability and statistics are intrinsically mathematical and symbol-based disciplines. In this author's opinion, there is no use in sugar coating this fact, attempting to minimize it, or avoiding it altogether. This textbook, therefore, uses mathematical symbols as necessary and demands that the ecosystem manager becomes well versed in the meanings of the symbols. Here, "demand" is operationalized by communicating exercises and their end point solution exclusively through mathematical symbols.

1.2.2.1 What can the reader hope to achieve?

An ecosystem manager who has completed all of the exercises in this textbook will *really* understand the following.

1. The difference between a random variable and a datum.
2. The difference between an estimator and an estimate.
3. Probability distributions, cumulative distribution functions (CDFs), probability density functions (PDFs), quantiles, and realizations (deviates).
4. How statistical inference serves ecosystem modeling.
5. How influence diagrams can be used to model ecosystems and the decision making of associated stakeholders.
6. How one can construct an IBM of an ecosystem process.
7. How to use stochastic models to reach effective but practical ecosystem management decisions.

1.2.3 Learning resources

1.2.3.1 Resources developed specifically for this textbook

JAVA source code files that constitute the **id** software system are available from the textbook's website. These files are sufficient to solve all exercises in the textbook. By preparing appropriate input files, the reader will be able to perform optimization, simulation, and resampling-based hypothesis tests. All exercise solutions are available on this website. Source code availability is needed in an introductory textbook on probability and statistics because algorithms operationalize probability and statistics concepts. Having access to the source code allows the ecosystem manager to see the details of an algorithm's functioning and hence fosters a deeper understanding of the algorithm being studied.

An OITS that is tailored to this textbook may be accessed for free. This system is built with Bayesian networks along the lines suggested by Nkambou, Mizoguchi, and Bourdeau (2010). This OITS allows readers who are self-studying this textbook to receive feedback on those content areas of the textbook that they do or do not understand.

1.2.3.2 R, R Commander, and quantum GIS

There are many excellent, large, and complicated statistical software systems such as SAS or SPSS. There are also many excellent, large, and complicated GIS packages such as ArcView. This textbook, although introductory, is focused on Bayesian networks, influence diagrams, simulation, resampling methods, spatial statistics, spatio-temporal statistics, and agent-based models. These areas are not well served by large statistics packages. Further, such large packages have steep learning curves, hide a statistical computation's intermediate steps, and can be expensive.

This textbook uses R, QGIS, and the **id** package exclusively in examples and for exercises so that all concepts and computations are visible and presented in a manner that emphasizes the central role of simulation, resampling methods, and the unifying modeling architecture of influence diagrams. QGIS is a free and quite extensive software system for analyzing spatial data (`http://qgis.org`). This textbook stresses the need to rely on high-performance, web-aware, and interpretable, open source code for statistical computations. A fundamental pedagogical precept of this textbook is that the understanding of a few statistical computations is essential for a beginner to develop a working and adaptable understanding of probabilistic models and statistical methods applied to ecosystem management.

To ease a beginner into the probability and statistics material, a free graphical user interface (GUI) statistics package with simulation capabilities is used in the chapters on introduction to probability and introduction to statistical inference (Chapters 2–5). This free GUI, called *R Commander*, is a front end to R (`http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/`). Two recommended introductory statistics texts that make use of R are by Dalgaard (2008) and Verzani (2004). An advanced text on linear models that contains R examples is that by Faraway (2006). Bivand, Pebesma, and Gómez-Rubio (2008) employ R to analyze spatial data.

To ease a beginner into the GIS material, QGIS is used to introduce enough GIS content to support the material on spatial and spatio-temporal statistics (Chapters 6 and 7).

In addition to possessing a GUI, both the introductory statistics package (here, R Commander) and the introductory GIS package (here QGIS) need to be free, easy to install, and easy to learn. The statistics package, in particular, needs to support simulation and have attractive plotting capabilities. R Commander and QGIS possess these characteristics.

Free packages are not supported indefinitely. In particular, should R be unavailable to the reader either now or in the future, the R codes presented herein have been written in such a way as to make them easy to translate into other computer languages that are capable of simulation.

Should it become necessary, R Commander could be replaced with SIMFIT (`www.simfit.manchester.ac.uk` or `www.simfit.silverfrost.com`), Instat (`www.reading.ac.uk/ssc/n/n_instat.htm#intro`), PSPP (`http://gnu.org/software/pspp/`), or the UNESCO package, WinIDAMS (`http://portal.unesco.org/new/en/communication-and-information/` then click "resources" followed by "software"). The R code examples in this textbook could be easily ported to SIMFIT codes. QGIS could be replaced with GRASS or JGRASS.

1.2.3.3 Additional resources

There are several good study guides on how to do well in a quantitative course. Two of these are Schiavone (1998) and Smith (1998).

There are several web-based resources for learning basic probability and statistics. Some of these are Kenny (1987), `www.psychstat.missouristate.edu/sbk00.htm`, and `www.sjsu.edu/faculty/gerstman/StatPrimer`

1.3 Chapter summaries

Chapter 2 Probability and simulation
This chapter explains basic frequentist probability ideas and definitions and introduces the use of software-based simulation with a simple example using R Commander; conditional probability is explained through the development of a simple Bayesian network. The chapter gives a detailed presentation of independent-deviate simulation only – Markov chain Monte Carlo algorithms are avoided in this textbook – and gives several examples.

Chapter 3 Application of probability: Models of political decision making in ecosystem management
This chapter briefly outlines the political–ecological system simulator of Haas (2011); it explains that the component models contained therein will be used as examples throughout the textbook. The chapter reviews alternative approaches to modeling the political process of reaching ecosystem management decisions and builds models of how two different groups reach ecosystem management decisions: wildlife management policymakers and rhino poachers.

Chapter 4 Statistical inference I: Basic ideas and parameter estimation
This chapter introduces statistical estimation of model parameters and explains how MSHD and its generalization, consistency analysis (see Haas (2011, ch. 4) and Chapter 4 herein), can be used to estimate the parameters

of any stochastic model. It describes the two essential elements of this parameter estimation method: the ability to generate realizations from the model and the ability to estimate a density nonparametrically, and gives an example using **id**. Note that this chapter on statistical estimation appears before the one on statistical inference (Chapter 5) as many hypothesis tests operate by comparing two estimated models.

Chapter 5 Statistical inference II: Hypothesis tests

This chapter describes the basics of a statistical hypothesis test, a confidence interval, and a prediction interval and shows how hypothesis tests can be obtained from confidence intervals. It describes a particular resampling-based test called the *delete-d jackknife hypothesis test* as a computer-intensive alternative to a test that requires an analytic form of the test statistic's distribution under the null hypothesis and gives examples of these tests and confidence intervals.

Chapter 6 Introduction to spatial statistics

This chapter gives a step-by-step guide to installing and using QGIS and introduces basic spatial statistical ideas such as the semivariogram, kriging, Ripley's K function for spatial point patterns, and spatial prediction as a conditional expected value. It gives a semi-parametric spatial predictor introduced by this author and applies it to sulfate deposition in the conterminous United States. The chapter presents the family of marked spatial point processes and introduces a multivariate spatial process and a moving window method for predicting its values.

Chapter 7 Introduction to spatio-temporal statistics

This chapter extends spatial statistics to the spatio-temporal case and describes spatio-temporal variables such as road density or vegetation index through time. It describes statistical algorithms to estimate such variables, describes how QGIS and **id** implement these algorithms, and gives an example of their use. There is a dearth of algorithms in available GIS software that can perform spatio-temporal statistical computations. Hence, this chapter includes a description of this author's approach to estimating the parameters of a semi-parametric spatio-temporal process model. An example of sulfate deposition through time across the conterminous United States is given. The spatio-temporal progress of an invasive bumblebee across Japan is given as an example of a spatio-temporal point process.

Chapter 8 Application of statistical inference: Estimating the parameters of an individual-based model

This chapter describes the basic structure of an IBM, explains how MSHD is used to estimate such a model's parameters, and gives two examples: bacterial growth on a biofilm and the rhino meta-population on private wildlife ranches in South Africa.

Chapter 9 Guiding an influence diagram's learning

This chapter briefly reviews Bayesian network learning algorithms. Specifically, a new learning algorithm is developed for the group decision-making models of Haas (2011) that allow ecosystem stakeholders to use the discrepancy between what they expected to happen and what actually happened to update their perceptions about likely consequences of their future decisions. The chapter gives an example of this algorithm's use.

Chapter 10 Fitting and testing a political–ecological simulator
This chapter pulls together all of the textbook's chapters to estimate the parameters of a political–ecological simulator of privately held South African rhino and finds the most practical ecosystem management plan (MPEMP) using this fitted simulator.

1.4 Installing and running R Commander

1.4.1 Running R

To begin an R session in Windows®, either double click the R icon or click the `Start` icon, then `all programs`, and then `R`. At the R prompt (>), set the working directory with `setwd` (*path*). For example, `setwd("c:/polbio/introtext")` sets the working directory to `c:/polbio/introtext`. Then, load all needed R packages by running the setup code for this textbook by typing `source("setup.r")` at the R command prompt. This file consists of the following.

```
#Filename: setup.r
require(compiler)
enableJIT(3)
#Packages may need to be installed first.  To do this with
#the "fields" package for example, issue the command:
#install.packages("fields")
library(mvtnorm)
library(fields)
library(spatstat)
library(maps)
library(mapproj)
library(sp)
library(gstat)
library(ads)
library(dfoptim)
library(np)
```

To redirect R output to a file, type the command `sink("out.txt",type=c('output'))`. To redirect the output back to the R console window, type the command `sink()`.

1.4.2 Starting an R Commander session

Once the R prompt is blinking, type the command `library(Rcmdr)`. If this command fails, you will need to first install the R Commander package with the command `install.packages("Rcmdr")`. To execute this command, your computer needs to be connected to the Internet. After the package is installed, retype the earlier `library` command to start an R Commander session.

1.4.3 Terminating an R Commander session

To end an R Commander session, click `File > Exit`.

1.5 Introductory R Commander session

You are monitoring fish uptake of toxic chemicals and you have recorded the data:

```
Chemical       Frequency
Mercury           92
PCBs              35
Dioxin/Furan       9
```

Do the following to display this data as a strip chart in R Commander.

1. Start an R session.
2. At the R prompt, type `library(Rcmdr)`.
3. In the R Commander window, click `Data > New data set`.
4. In the `Enter name for data set`: dialog box, type `fish` and then click `OK`.
5. Type the above data on toxic chemicals in fish into the data spreadsheet. If you make a mistake, make the cell editable by double clicking it.
6. Click the red `x` to close the spreadsheet window and load the data into R Commander's *active* dataset.
7. Click `Graphs > Strip chart > OK`.

You should end up with the R Commander window as shown in Figure 1.1, and the strip chart as in Figure 1.2.

Example 1.1 Summary statistics on a dataset may be computed in R Commander with the following sequence of menu selections and inputs.

1. Click `Data > New data set`.
2. Type `onetoten` for the dataset's name.
3. Enter the numbers 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10 into column 1 of the spreadsheet.
4. Click the red `x` in the upper-right corner of the spreadsheet's window.
5. Click `Statistics > Summaries > Active data set`.

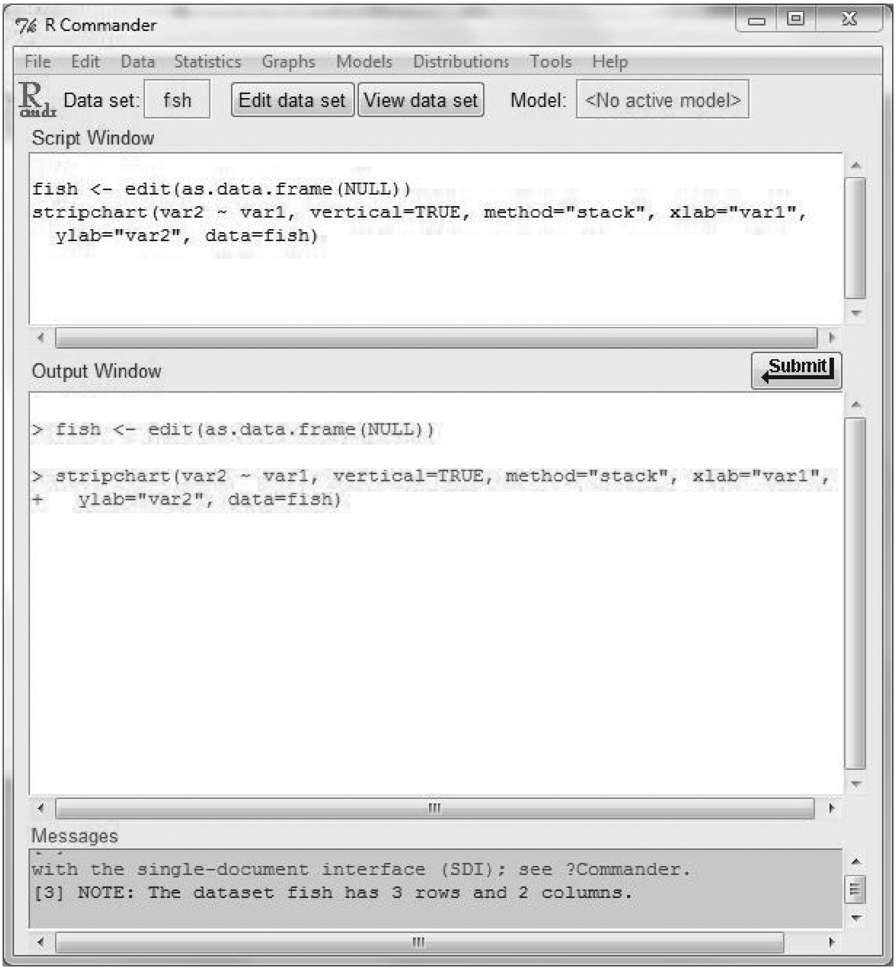

Figure 1.1 R Commander's window just after the fish toxic chemical data strip chart has been drawn.

Example 1.2 To issue an R command within the R Commander, type the commands to be executed into the Script Window. Then, after highlighting them, click the `Submit` button to the right of the R Commander screen. For example, to obtain a stem-and-leaf diagram of the data entered in Example 1.1, do the following.

1. Click anywhere in the Script Window. Then, type in the following two lines:

```
onetoten
stem(onetoten$var1)
```

2. Place the mouse on the first character of the first of these lines. Hold down the left mouse button and drag the mouse down until both of these lines

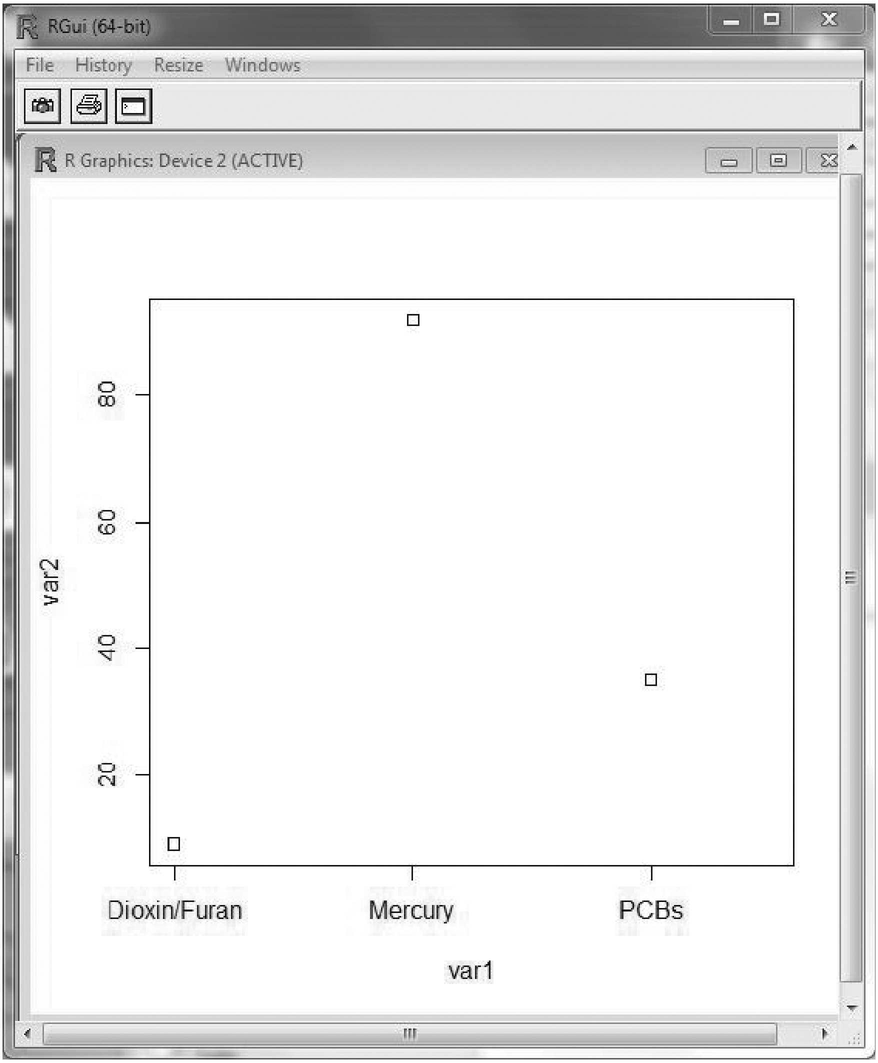

Figure 1.2 Strip chart of fish toxic chemical data as drawn by R Commander.

are highlighted in blue. The mouse cursor looks like an uppercase letter "I." Some authors refer to the mouse cursor as the *editing I beam*.

3. Release the left button on the mouse. Then, move the mouse to the `Submit` button and left click it.

Your screen should be similar to that of Figure 1.3.

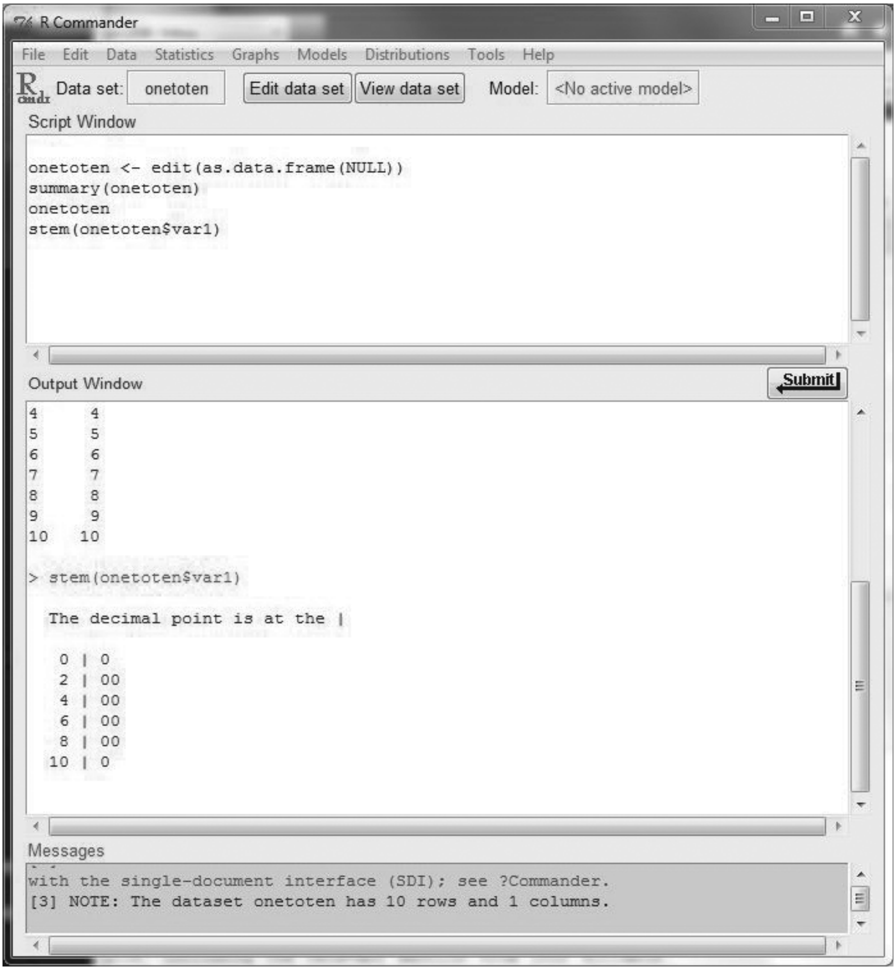

Figure 1.3 Entering R commands from R Commander's Script Window.

1.6 Teaching probability through simulation

There are many excellent introductory probability textbooks that take an axiomatic approach. One modern and well-written exemplar that avoids calculus (for the most part) is by Santos (2011).

There is a dearth of texts, however, that approach the material through simulation. Simulation mimics the physical reality that probability describes. These underemphasized connections between probability, physical mechanisms, and simulation, have led Yakowitz (1977, pp. xvi) to quote two eminent probabilists:

> In fact, all epistemologic value of the theory of probability is based on this: that large-scale random phenomena in their collective action create strict non-random regularity. The very concept of mathematical probability would be fruitless if it did not find its realization in the frequency of occurrence of events under large-scale repetitions of uniform conditions (a realization which is always approximate and not wholly reliable, but that becomes, in principle, arbitrarily precise and reliable as the number of repetitions increases) (Gnedenko and Kolmogorov, 1954).

Yakowitz (1977, pp. xvi–xvii) continues by noting that his experience with students suggests that studying derivations of probability distributions and proofs of limit theorems does little to motivate or give insight into a distribution's character or a theorem's statement.

Yakowitz (1977, pp. xvii) states that in his textbook, simulation experiments are set up so that the student may see in real, physical terms the behavior described by the mathematical form of a distribution or the statement of a theorem. One well-known use of this approach to demonstration is the use of increasingly large sample sizes drawn from some highly non-normal distribution to illustrate the meaning of the central limit theorem.

The present textbook takes the idea of a simulation-based approach to probability one step closer to probability's physical roots by replacing mathematical derivations as much as possible with simulation algorithms. This approach is especially evident with regard to the counting rules used to explain discrete probability distributions such as the binomial.

This emphasis on simulation is not an attempt to discount the mathematical underpinnings of probability but rather to (i) make probability more approachable to those with modest mathematical backgrounds, (ii) emphasize and promote the physical and mechanical nature of many probability distributions, and (iii) deliver an intuitive explanation of probability ideas in a compact and immediately understandable way.

1.6.1 The frequentist statistical inference paradigm

There are two main approaches to statistical inference: *frequentist* and *Bayesian*. Both approaches view probability distribution parameters (Chapter 2) as having fixed but unknown values. Also, both approaches view the goal of statistical inference to be the finding of some form of a numerical estimate of these true values based on a set of observations. A Bayesian approach to this task is to first have the ecosystem manager express his or her pre-observation beliefs about a parameter's value in the form of a probability distribution. Doing so makes the parameter a random variable (Chapter 2) during the course of the statistical inference computation and subsequent interpretation of the computation's result. A frequentist-based statistical inference procedure does not require the ecosystem manager to express his/her pre-observation beliefs in the value of the parameter and consequently does not represent that parameter as a random variable.

This text restricts itself to the frequentist paradigm of statistics in its presentation of statistical model estimation, statistical inference, and learning algorithms for influence diagrams.

1.7 Summary

This chapter has set the stage for what a reader can expect as he or she works through the textbook's chapters. Justifications have been given for why the text's content is crucial to the practice of modern ecosystem management and for why probability ideas will be developed through the vehicle of simulation. Statistical and GIS software packages that the text uses have been introduced. The reader will learn the text's material more easily if he or she takes the time to run each computational example's R code, **id** input file, or sequence of GUI steps within QGIS.