A Primer in Biological Data Analysis and Visualization Using R

Gregg Hartvigsen

A Primer in Biological Data Analysis and Visualization Using R

Gregg Hartvigsen

COLUMBIA UNIVERSITY PRESS NEW YORK

Columbia University Press
Publishers Since 1893
New York Chichester, West Sussex
cup.columbia.edu

Library of Congress Cataloging-in-Publication Data

Hartvigsen, Gregg.
A primer in biological data analysis and visualization using R / Gregg Hartvigsen
p. cm.
Includes [bibliographical references and index.]
ISBN 978-0-231-16698-0 (cloth : alk. paper) — ISBN 978-0-231-16699-7 (pbk. : alk. paper) — ISBN 978-0-231-53704-9 (e-book)

Library of Congress Subject Data and Holding Information can be found on the Library of Congress Online Catalog.

2013952140

Columbia University Press books are printed on permanent and durable acid-free paper.
This book is printed on paper with recycled content.
Printed in the United States of America

c 10 9 8 7 6 5 4 3 2 1
p 10 9 8 7 6 5 4 3 2 1

Cover design: Milenda Nan Ok Lee
Cover image: © Getty Images

References to websites (URLs) were accurate at the time of writing.
Neither the author nor Columbia University Press is responsible for URLs that may have expired or changed since the manuscript was prepared.

A PRIMER IN BIOLOGICAL DATA ANALYSIS AND VISUALIZATION USING R

>>>>>>>>>>>>>>>>>>>>>>>>

A PRIMER IN BIOLOGICAL DATA ANALYSIS AND VISUALIZATION USING R

>>>>>>>>>>>><<<<<<<<<<<<

CONTENTS

INTRODUCTION

> We face danger whenever information growth outpaces our understanding of how to process it.
>
> (Silver, 2012)

In our effort to understand and predict patterns and processes in biology we usually develop an idea or, more formally, a conceptual model of how our system works. We generally frame our models as testable hypotheses that we challenge with data. As the science of biology has matured our questions of how nature works have gotten more sophisticated and complex. Unfortunately, we are not able to simply look at a table of raw data that we get from an experiment and see an answer to an interesting question with any quantitative level of confidence. Instead, to accomplish this we will learn how to use the R statistical and programming software package to process these data (summarize, analyze, and visualize our results). We also will go a step further and work to understand what these results mean biologically.

Data, graphs, and statistics, oh my! Isn't the interesting stuff in biology really just the cool, living things all around us? It is that stuff but it's *so much more beautiful* when we understand it. Maybe you want to be a vet. Perhaps an early memory for you was loving a little furry thing that purred. However, maybe now you've become a little more concerned about what impact these lovable pets might have on populations of other cute animals that live outside. I recently took a break from writing and looked at an issue of the journal *PLoS ONE* (a well-respected, open-access, online journal). In this journal I saw an article on predation by urban cats in the UK (Thomas et al. (2012)). I "own" three cats and was surprised by the number of prey items that cats brought back to their owners (see Figure 1). It seems that there is a lot of variability

in predation rates (the histogram) and that predation rates decrease with increasing urbanization (housing density). Specifically, as seen in the inset graph, the authors state that "There was a significant negative correlation between housing density and annual predation rates on birds ($r = 20.699$, $p = 0.036$)."

When we have questions that we want to answer, such as "what are cats up to when they're outside?," we might read books of fiction, such as the series on Warrior cats (see books by Erin Hunter, which is actually a pseudonym!). In biology, however, we seek to understand things like cats by collecting, interpreting, analyzing, and visualizing data. This book is designed to help you to be able to do this. If you're interested in other disciplines I hope the examples in this book help you, too! I also hope that as you use this book you lose any fear you might have of data and instead seek out and work with data and understand what they tell you about the things that got you interested in biology in the first place, like cats (or, more likely, dogs).

WHAT THIS BOOK IS (AND ISN'T)

This book is designed to help you collect, organize, analyze, and visualize data. I assume you have not heard of the free, open-source program R and I will, therefore, introduce you to how to use it to accomplish these goals. Although I imagine you have had some experience making graphs and calculating a few descriptive statistics (e.g., mean and standard deviation in Excel) I assume you haven't done this. If you don't know Excel, or don't have access to it, you will be able to do all the heavy lifting in this book. I assume you have not taken a course in statistics.

This book, therefore, aims to give you a foundation upon which to become a better student of science and a better consumer of scientific information. More specifically you will learn how to

- formulate hypotheses,
- design better experiments,
- do many standard statistical procedures,
- interpret your results,
- create publication-quality visualizations of your results,
- find help so you can solve your own problems, and
- write a simple computer program.

You shouldn't expect to read this book and become a quantitative guru. Instead, you should hope to become competent at finding answers to some

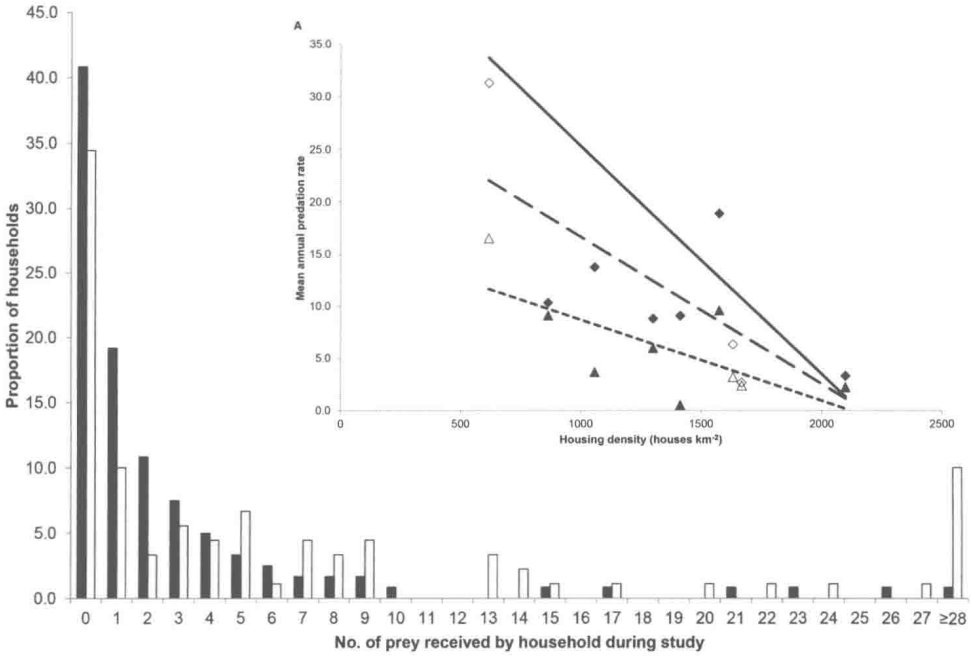

Figure 1: Two figures from a recent paper on urban cat predation rates (Thomas et al. [2012]). The larger graph is a histogram showing percentages (instead of the usual frequencies, or counts) for the number of prey returned to households. Black and white bars are for households with a single-cat versus multiple-cats, respectively. The insert is a scatterplot with best-fit straight lines added for birds, mammals, and for both animal groups combined. The combined data points have been omitted! The relationships are analyzed and discussed in the paper as "correlations" and, therefore, adding lines is inappropriate (see the box on page 138). The graphs and resulting analyses were likely done using R, but that doesn't mean they are correct! After you work through this introduction you should be able to comfortably assess these data, correctly perform the analyses and create more appropriate visualizations.

of your questions, such as "are these two samples different?" and "is there a significant linear relationship between my variables?" You will become a resource to the people around you. And if you put in some time playing with R you will be the go-to person for data.

I have written this book primarily with the hope that you'll feel more comfortable with complex biological problems. It has grown out of what I

have seen challenge my own undergraduate students. But it also covers some topics that I think are fun and valuable to know how to do (e.g., programming). The chapters end with problem sets for you to challenge yourself to use what you have learned. Some of the data are real while some are merely *realistic*. I also have included solutions to the odd-numbered problems at the end of the book. Finally, the book is filled with R code. You should type this is in yourself because this helps with the learning process. You can, however, go to `https://github.com/GreggHartvigsen/PrimerBiostats` and download all the code from this book.

This book is neither a formal introduction to R nor a statistics textbook. Instead, this book helps you to you solve problems you're likely to encounter in your undergraduate program in biology. I work to explain what statistics are and how to share and interpret scientific results. After working through this book you should be able to solve a variety of problems with the most widely used statistical and programming environment. I hope you will no longer be afraid of data and will be more able to enter data into the computer, test hypotheses, and present your findings.

So, this book should help you make more appropriate and professional, scientific visualizations and discover findings that might have otherwise been missed. You will no longer be satisfied with hearing from anyone things like "Well, it looks significant" or "there seems to be a trend in the data." So, for the rest of your career, I hope you become the person who says "We can test that! Let me get my laptop."

WHO REALLY NEEDS THIS?

In this book I work not only to present visualization and analytical techniques but to explain why we do all this. There's an unfortunate misconception that we don't really need all this quantitative stuff in biology. I have heard several times the following line of thinking:

> Why do we need to use statistics in biology? If the hypothesis is clear, the experiment is designed correctly, and the data are carefully collected, anyone should be able to just look at the data and clearly see whether or not the hypothesis is supported. Statistical procedures are simply safety nets for sloppy science.

As you work your way through this book you'll see why the above thinking limits scientific exploration, understanding, and the ability to make predictions