Synthesis Lectures on
Mathematics & Statistics

SYNTHESIS
COLLECTION OF TECHNOLOGY

Robert Hirsch

Analysis of Epidemiologic Data Using R

Springer

Synthesis Lectures on Mathematics & Statistics

This series includes titles in applied mathematics and statistics for cross-disciplinary STEM professionals, educators, researchers, and students. The series focuses on new and traditional techniques to develop mathematical knowledge and skills, an understanding of core mathematical reasoning, and the ability to utilize data in specific applications.

Robert Hirsch

Analysis of Epidemiologic Data Using R

Springer

Robert Hirsch
Overland Park, KS, USA

ISSN 1938-1743 ISSN 1938-1751 (electronic)
Synthesis Lectures on Mathematics & Statistics
ISBN 978-3-031-41913-3 ISBN 978-3-031-41914-0 (eBook)
https://doi.org/10.1007/978-3-031-41914-0

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Paper in this product is recyclable.

Preface

I had two goals in writing this text. One was to bring together methods of analysis for epidemiologic data in a way that did not require mathematics to understand the principles. This was accomplished with the only math being the methods themselves. The second goal was to make these methods accessible using R. This also was accomplished limited only by the existence of programs.

In addition to chapters addressing the statistical methods, there are four chapters addressing related topics. First, I assume that the reader has no familiarity with R, so there is a chapter that describes how to access R and how to do the types of things required to perform an analysis of epidemiologic data. Next, I assume that the reader has little familiarity with statistical methods, so there is a chapter describing the statistical logic behind estimation and hypothesis testing. Next, I assume that the reader has little familiarity with causal inference, so there is a chapter describing the concept of causation and how statistical methods can be used to support a causal conclusion. Finally, I assume that the reader has little familiarity with epidemiologic research designs, so there is a chapter describing the most common experimental and observational studies.

Where my assumptions about the reader are incorrect, the chapters that discuss topics with which the reader is familiar can be skipped without sacrificing appreciation of the remaining five chapters. In essence, the experienced reader can start with Chap. 5.

In selecting the topics of the last five chapters, I tried to cover all of the most common methods to analyze epidemiologic data. These chapters begin by discussing measures of disease frequency and measures of association using 2×2 tables as the structure. Then, I discuss life table analysis as a method for interpreting incidence data as risks. Next, I address stratified analysis beginning with an introduction to confounding. Finally, I present regression analysis as another method to control for confounding.

Overland Park, USA — Robert Hirsch

Acknowledgments I appreciate my clients who have shared their frustration with trying to understand statistical methods for epidemiologic data and their desire to be able to perform some statistical procedures without having to involve a statistician.

Notices

The examples in this book are based on fictitious data and should not be taken as a reflection of real relationships. Those data have been created only to illustrate statistical principles.

Contents

Introduction to R

1

Abstract

R is free statistical software that is available for download on the web. It is usually run from an interface called "RStudio." RStudio is set up with four quadrants. The lower left is where the programs are run. The lower right is where the documentation is available and where graphic output appears. The upper right is where datasets are listed, and the upper left is where you can see the content of datasets. Creation of datasets usually starts with the creation of vectors in the lower left quadrant. These vectors can be combined to create a dataset or data frame. A special kind of dataset is stored in a 2×2 matrix. These are often of use in epidemiology. We will be using several programs. These are arranged in packages. The "stats" package is the one we will use most often. Click on the name in the lower right quadrant to see a list of the programs available. Clicking on the name of the program invokes documentation on that program. Graphics are requested in the lower left quadrant, and they appear in the lower right quadrant. We use the "plot" command to create graphs. The plot command has several options used to customize your graph. The graphs can have additional data displayed using the "points" command. You can create a legend for your graph using the "legend" command. A special kind of graph used in epidemiology can be created using the "stepfun" function.

R is a collection of programs for statistical analysis and data management. It was first developed by two New Zealand statisticians, Ross Ihaka and Robert Gentleman and released in 1993.[1] It has since been expanded and modified by the R Core Team and the

[1] R got its name from the first letter of the developers' first names.

R. Hirsch, *Analysis of Epidemiologic Data Using R*, Synthesis Lectures on Mathematics & Statistics, https://doi.org/10.1007/978-3-031-41914-0_1