

# Introduction to Data Science in Biostatistics

Thomas W. MacFarland

# Introduction to Data Science in Biostatistics

Using R, the Tidyverse Ecosystem, and APIs

 Springer

Thomas W. MacFarland  
Office of Institutional Effectiveness and College  
of Computing and Engineering  
Nova Southeastern University  
Fort Lauderdale, FL, USA

ISBN 978-3-031-46382-2      ISBN 978-3-031-46383-9 (eBook)  
<https://doi.org/10.1007/978-3-031-46383-9>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2024

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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.

*Airson mo theaghlach agus am foighidinn  
gun chrìoch.*

# Foreword

Beginning with the mid-1970s development of S and its reimagination into R, approximately 20 years later, R remains a leading language in biostatistics. By the mid-2000s, ease of use and functionality with the R language expanded greatly when the tidyverse ecosystem saw its first implementation.

Throughout its evolution to today, R has remained open-source software that is freely available to all. From among its many uses, R supports data acquisition from distant hosts using Application Programming Interface (API) clients, data management and data organization using tidyverse ecosystem tools such as the `dplyr` package and the `tidyr` package, and superior production of graphics and maps using the ubiquitous tidyverse `ggplot2` package and complementary packages that are `ggplot2` compliant. There is also a host of other R-based tools for statistical analyses that *work and play well* with APIs and the pervasive tidyverse ecosystem. It is argued in this text that R should always be among the first selections in any list of software that supports biostatistics.

This text was developed to assist beginning students and early stages researchers in their attempt to make sense of how software can be used in biostatistics, viewing an all-pervasive concept of biostatistics in the large and the many disciplines associated with biostatistics. To meet this challenge, R was selected as the most appropriate programming language, calling on Base R (e.g., the many functions made available when R is first downloaded) and supporting packages (e.g., the thousands of auxiliary R software collections that provide functionality far beyond what is available in Base R, especially packages associated with the tidyverse ecosystem).

The desire to prepare an introductory text that was based on the needs of beginning students and early stages researchers grew out of observations from prior teaching experiences in biostatistics, individually and of colleagues. It could not be ignored that frustrations of those who were new to biostatistics were so great that many who might have later taken on careers and leadership roles in biostatistics moved on to other fields of study, greatly impacting the immediate need for favorable retention rates and future concern about lost intellectual potential and leadership for the profession. It is hoped that the many details and examples shown in this text, admittedly verbose for those with experience but needed for the target

audience of this text, will help those who are still new at biostatistics and in turn improve retention of students and career advancement of future biostatisticians.

Data scientists provide value beyond the immediate. Following along with this concept, value is added to this text in that most lessons are enhanced by greatly detailed addenda, often multiple addenda in each lesson. New ideas, exposure to new tidyverse ecosystem packages and functions, and needed skills are gradually addressed with each advancing lesson. The addenda often introduce, reinforce, and expand on specialized tidyverse ecosystem packages and functions that go beyond what was previously presented, address parametric versus nonparametric approaches toward data, and often include practice data sets that support incremental engagement in pursuit of advanced skills.

Of equal importance to those with interest beyond any cursory introduction to biostatistics, many challenge activities are included throughout each lesson and the addenda. The challenges at first are simple and should be successfully completed by all. Later, as the text continues, the challenges are more detailed, calling for creative attempts to achieve success, with some challenge activities lacking complete guidance, purposely. The later challenges are often worded so that there is no one and only one correct approach to resolution but instead the challenges allow for multiple approaches at resolution. A few of the last challenges also call for individual inquiry into more advanced topics and resources in the use of R with biostatistics than what is presented in the text. Not to be redundant, but these later challenges will indeed be challenging, but of course data scientists face challenges daily.

Additional value is also added in that each external dataset mentioned and used has been placed at the publisher's Web site for this text. These datasets are easily available for download, and their inclusion makes it possible to follow along with the syntax presented in the text. Ideally, use the syntax and provided datasets to reproduce the outcomes shown in this text. Then, go beyond the original syntax and try different approaches to data organization, experiment with other data analysis approaches, and consider additional functions and function arguments to produce even more enhanced figures and maps, etc. Take on the role of a data scientist and add value beyond base requirements.

Dr. Thomas W. MacFarland  
Senior Research Associate  
Office of Institutional Effectiveness; and Associate Professor  
College of Computing and Engineering  
Nova Southeastern University  
3301 College Avenue  
Fort Lauderdale  
Florida 33314

# Preface

The use of R and specifically the use of APIs and R's evolving tidyverse ecosystem for engagement in biostatistics is the focus of this text. By following along with a gradual exposure to R, APIs, and the tidyverse ecosystem, this text should help beginning students and early stages researchers gradually increase their skills with the use of R syntax for inquiries into biostatistics.

The first lesson of this text is somewhat unique in that it looks closely at the way data science is viewed as an emerged (not emerging) discipline in higher education. The United States Department of Education has a hierarchical coding system for the way academic majors are organized, and from this system, a large collection of majors that call for some degree of expertise in data science is identified. These majors are then put into context by the hierarchical coding system used by the United States Department of Labor and the eventual transition from academic preparation to careers. Although higher education has experienced a noticeable decline in enrollment over the last few years, that is not the case for data science. There is clearly an increase in interest in data science, not surprisingly due to the growth of data science as a career opportunity. Employment in data science is projected to grow and salaries are projected to increase. A few basic summaries on the use of R and data types are also stressed in the first lesson, as either a recapitulation for those with prior exposure to R or as an introduction for those who are not as well versed in the use of R and how data are viewed.

The next two lessons look closely at data. A summary of possible data sources related to biostatistics is the focus of the second lesson. Although it may seem intuitive to those with experience, beginning students and early stages researchers need to know that there are many resources that either provide data that may totally meet needs as inquiries are attempted, or the data may serve as a useful proxy. Government data sources are especially valued and are stressed in this lesson. Knowing possible data resources, the third lesson stresses a curated ten-point process at statistical analyses, with an emphasis in the lesson on how these processes are used with an all-inclusive demonstration of statistical tests.

The process stressed in the third lesson leads to a more detailed introduction to the tidyverse ecosystem in the fourth lesson. Emphasis is placed on how the

tidyverse ecosystem is used to organize workflow, as inquiries into biostatistics are attempted. The fourth lesson goes into detail on core tidyverse ecosystem packages and auxiliary packages that complement a tidy workflow. These R packages and their many associated functions and arguments are then detailed throughout the remaining parts of this text.

The fifth lesson is focused on statistical analyses. Specific tests are demonstrated and there is also considerable discussion on the issue of parametric versus nonparametric approaches to statistical testing.

In contrast to the use of a GUI (Graphical User Interface) and click-type selections to build and download a dataset, the sixth lesson emphasizes an API (Application Programming Interface) approach to data acquisition. An API consists of syntax within an R work session and the use of syntax is a far more efficient and reproducible means of obtaining data than undocumented GUI selections. Different resources that support R-based APIs are identified in the sixth lesson.

The concluding seventh lesson provides a detailed summary of what was covered throughout the text, including: how data are obtained using an API; how data are put into tidy format; how data are subjected to statistical tests; and how data are used to create a wide variety of figures, including maps. Along with use of the data, a few ideas on how data scientists prepare summary reports are also demonstrated. The ending lesson is worded to look forward to the world of data science and how R is used to support inquiries, with ending comments on the favorable future of data science, along with general ideas about professionalism and soft skills in the data science workplace as well as data science in the large. Finally, the text ends with information needed to contact the author and a reminder on how to obtain all datasets referenced in this text.

Fort Lauderdale, FL, USA  
Fall 2023

Thomas W. MacFarland



# Acknowledgments

I cannot begin to adequately thank the many individuals who contribute to the open-source paradigm and the countless number of hours given freely to software development and management, often for little if any financial remuneration and only rare acknowledgment by deans and supervisors as a metric for career advancement. These individuals put the profession and the advancement of science first, often at the cost of time away from personal pursuits.

I also want to recognize all at Springer who assisted with this text, editors Laura Aileen Briskman and Faith Su and the many staff members, domestic and foreign, who make final production of disparate files into a cohesive text. To all – thank you for your many ideas, feedback, help, and supporting efforts.

# Contents

<b>1</b>	<b>Emergence of Data Science as a Critical Discipline in Biostatistics</b>	<b>1</b>
	Definition and History of Data Science	1
	The State of Data Science and the Need for Data Scientists	3
	Definition of Data	3
	Emergence of Data as a Valued Problem-Solving Input.	6
	Emergence of Data Science as a Highly Valued Occupation and Career Paths	9
	Biostatistics: Definition and Applications Allowing for Frequent Overlap	9
	Academic Growth of Data Science Programs of Study in the Biological Sciences, Based on Classification of Instructional Programs (CIP) Codes Related to Biostatistics	11
	CIP Series 01: Agricultural, Animal, Plant, Veterinary Science and Related Fields.	14
	CIP Series 26: Biological and Biomedical Sciences.	14
	CIP Series 27: Mathematics and Statistics	14
	CIP Series 30: Multi-Interdisciplinary Studies.	15
	CIP Series 44: Public Administration and Social Service	15
	CIP Series: 51: Health Professions and Related Programs.	15
	Jobs and Job Requirements for a Data Scientist.	16
	Job Opportunities and Salaries in Data Science	16
	Job Opportunities and Salaries in Data Science	20
	Computing and Data Science	26
	Pre-ENIAC (1946)	26
	Mainframe Computing (1950s Onward).	27
	Personal Computing (1980s Onward).	27
	Widespread Acceptance of the Internet (1970s Onward) and the World Wide Web (1989 Onward).	28
	Movement to Cloud Computing (2006 Onward)	29
	Data Types Supported by R	30

- Boolean (e.g., Logical) Data Expressing Comparisons and Order of Operations . . . . . 32
- Numeric Data . . . . . 38
- String or Character Data . . . . . 43
- Time and Dates . . . . . 47
- Missing Data . . . . . 54
- Data Structures Used in R . . . . . 60
- Addendum 1: Syntax Used to Generate Six-Digit Classification of Instructional Programs (CIP) Completions . . . . . 65
- Addendum 2: National and State Data for OCC-Identified Jobs Associated with Data Science and Biostatistics . . . . . 87
- External Data and/or Data Resources Used in This Lesson . . . . . 99
- 2 Data Sources in Biostatistics . . . . . 101**
  - Personal Data Sources . . . . . 101
  - Local Data Sources . . . . . 102
  - State Data Sources . . . . . 104
  - National Data Sources . . . . . 105
    - United States Census Bureau . . . . . 106
    - United States Centers for Disease Control and Prevention . . . . . 106
    - United States Department of Agriculture . . . . . 107
    - United States Department of Education . . . . . 109
    - United States Department of Labor . . . . . 110
    - United States Environmental Protection Agency . . . . . 111
    - United States National Science Foundation . . . . . 112
  - International Data Sources . . . . . 112
    - European Centre for Disease Prevention and Control . . . . . 112
    - The Organization for Economic Co-operation and Development . . . . . 113
    - Our World in Data . . . . . 114
    - United Nations Food and Agriculture Organization . . . . . 114
    - World Bank . . . . . 115
    - World Health Organization . . . . . 115
  - Proprietary and Other Resources . . . . . 116
    - Google Cloud Platform Datasets for COVID-19 Research . . . . . 117
    - New York Times COVID-19 Data at github . . . . . 117
  - Addendum 1: Our World in Data . . . . . 118
  - Addendum 2: United States Department of Labor, Bureau of Labor Statistics . . . . . 132
  - External Data and/or Data Resources Used in This Lesson . . . . . 144
- 3 Role of Statistics for Decision-Making in Biostatistics . . . . . 147**
  - Ten-Point Process When Using R for Statistical Analysis . . . . . 147
    - Identify Problems That Benefit from Statistical Analysis . . . . . 147
    - Identify Potential Data Resources . . . . . 148
    - Obtain the Data . . . . . 149

- Identify and Organize the Data and All Relevant Variables . . . . . 149
- Outline Potential Approach(s) for Analyses and Consider Alternate Approaches . . . . . 150
- Put Plans into Action, with Frequent Checks for Quality Assurance . . . . . 150
- Individual Review of All Outcomes . . . . . 150
- External Review of Outcomes Whenever Possible . . . . . 151
- Report at an Appropriate Level for the Intended Audience . . . . . 151
- Debrief to Establish Processes for Future Improvements. . . . . 151
- General Approach When Using R for Statistical Analysis . . . . . 152
  - Exploratory Graphics . . . . . 152
  - Exploratory Descriptive Statistics and Measures of Central Tendency . . . . . 152
  - Exploratory Analyses . . . . . 153
- Addendum: Use Inferential Statistics and R Syntax to Address Differences in Percentage Deaths from COVID-19 by the Urban v Rural Continuum. . . . . 153
- External Data and/or Data Resources Used in This Lesson . . . . . 173
- 4 Data Science and R, Base R, and the tidyverse Ecosystem . . . . . 175**
  - Workflow for Reproducible, Efficient, and Accurate Analyses and Presentations . . . . . 175
  - Base R . . . . . 179
  - The tidyverse Ecosystem . . . . . 181
    - The tidyverse Ecosystem as an Idea and the Need for Tidy Data. . . . . 182
    - The Core tidyverse Ecosystem as a Set of Tools in R Packages for Data Science . . . . . 184
    - Auxiliary Packages Outside of the Core tidyverse Ecosystem. . . . . 185
  - Addendum 1: Complex Data Set on Birth Rates Easily Accommodated by Using the tidyverse Ecosystem . . . . . 187
  - Addendum 2: Complex Data Set on Gross Domestic Product (GDP) and Comparison to Birth Rates by Using the tidyverse Ecosystem . . . . . 206
  - Addendum 3: Individual Initiative of Planned Workflow, Analyses, and Graphical Presentations . . . . . 213
  - Addendum 4: Essential tidyverse Ecosystem Functions That Every Data Scientists Should Master . . . . . 217
  - External Data and/or Data Resources Used in This Lesson . . . . . 219
- 5 Statistical Analyses and Graphical Presentations in Biostatistics Using Base R and the tidyverse Ecosystem. . . . . 221**
  - Overview of Using R for Statistical Analysis . . . . . 221
    - Background. . . . . 221
    - Import Data . . . . . 222
    - Code Book and Data Organization . . . . . 223

Exploratory Graphics . . . . .	223
Exploratory Descriptive Statistics and Measures of Central Tendency . . . . .	224
Exploratory Analyses . . . . .	224
Presentation of Outcomes . . . . .	225
Examples of Leading Statistical Tests, Including All Syntax and Presentation of Screen Outcomes and Graphics . . . . .	225
Nonparametric Tests . . . . .	225
Parametric Tests . . . . .	226
Addendum 1: A Parametric Approach to Statistical Analyses and Graphical Presentations for Data on Rates of Births and Rates of Deaths . . . . .	227
Background . . . . .	233
Import Data . . . . .	235
Code Book and Data Organization . . . . .	236
Exploratory Graphics . . . . .	237
Exploratory Analyses . . . . .	263
Presentation of Outcomes . . . . .	295
Addendum 2: A Nonparametric Approach to Statistical Analyses and Graphical Presentations for Data on Rates of Births and Rates of Deaths . . . . .	296
Addendum 3: Data Wrangling, and Then Statistical Analyses and Mapping . . . . .	301
Background . . . . .	302
Import the Data . . . . .	302
Code Book and Data Organization . . . . .	308
Exploratory Graphics . . . . .	309
Exploratory Descriptive Statistics and Measures of Central Tendency . . . . .	310
Exploratory Analyses . . . . .	310
Presentation of Outcomes . . . . .	311
Addendum 4: Prediction . . . . .	312
Background . . . . .	312
Code Book . . . . .	313
Import the Data . . . . .	314
Graphics (e.g., Figures and/or Maps) . . . . .	316
Exploratory Descriptive Statistics and Measures of Central Tendency . . . . .	318
Exploratory Analyses . . . . .	322
External Data and/or Data Resources Used in This Lesson . . . . .	339
<b>6 Use of R-Based APIs (Application Programming Interface) to Obtain Data . . . . .</b>	<b>341</b>
Emergence of APIs as a Data Resource . . . . .	341
APIs and Reproducible Syntax . . . . .	342

- APIs and the Need for a Key ..... 343
- Structure of an API to Automate Data Retrieval..... 345
- Structure of Data Returned by an API ..... 355
  - Data in Returned Format..... 355
  - Data After Organization and Manipulation with Tidyverse  
Tools ..... 356
- Common API Resources in Biostatistics, Government  
and Proprietary ..... 362
- Addendum 1: Use of the tidyUSDA::getQuickstat() API..... 363
- Addendum 2: Use an API to Obtain Multiple Files, Wrangle  
the Data, Merge Files, Review Absolute and Percentage Change  
Over Time ..... 388
  - Obtain Data on Iowa Corn Prices, 1867 Onward ..... 389
  - Obtain Data on Iowa Corn Acreage, 1926 Onward ..... 392
  - Wrangle the Data into a Singular Dataset..... 395
- Addendum 3: Use of Known URLs as a Proxy API (Application  
Programming Interface) ..... 403
- Addendum 4: API-Based Data in JavaScript Object Notation  
(JSON) Format ..... 424
- External Data and/or Data Resources Used in this Lesson..... 430
- 7 Putting It All Together – R, the tidyverse Ecosystem, and APIs. . . .** 433
  - Obtain Data from an API ..... 433
  - Make the Data Tidy..... 449
  - Statistical Tests – Base R and tidyverse Ecosystem Functions..... 459
  - Beautiful Graphics ..... 472
    - Grouped Data ..... 472
    - Interval and Real Numeric Data ..... 472
    - Beautiful Maps ..... 473
  - R Markdown and LaTeX Demonstrations of a Summary  
Memorandum of Findings..... 519
    - R Markdown ..... 520
    - LaTeX ..... 521
  - Concluding Comments and Next Steps ..... 521
    - Technical Skills of a Data Scientist ..... 522
    - Soft Skills of a Data Scientist..... 522
    - Future Employment Opportunities ..... 523
  - Contact the Author ..... 523
  - External Data and/or Data Resources Used in This Lesson ..... 523
- Index..... 525**