

LEARNING MADE EASY



# R

ALL-IN-ONE

for  
**dummies**<sup>®</sup>  
A Wiley Brand



**5**  
**Books**  
in one!

**Joseph Schmuller, PhD**

Author of *Statistical Analysis with R  
For Dummies*





# R

ALL-IN-ONE

by Joseph Schmuller

for  
**dummies**<sup>®</sup>  
A Wiley Brand

## R All-in-One For Dummies®

Published by: **John Wiley & Sons, Inc.**, 111 River Street, Hoboken, NJ 07030-5774, [www.wiley.com](http://www.wiley.com)

Copyright © 2023 by John Wiley & Sons, Inc., Hoboken, New Jersey

Published simultaneously in Canada

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without the prior written permission of the Publisher. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permissions>.

**Trademarks:** Wiley, For Dummies, the Dummies Man logo, Dummies.com, Making Everything Easier, and related trade dress are trademarks or registered trademarks of John Wiley & Sons, Inc. and may not be used without written permission. All other trademarks are the property of their respective owners. John Wiley & Sons, Inc. is not associated with any product or vendor mentioned in this book.

LIMIT OF LIABILITY/DISCLAIMER OF WARRANTY: WHILE THE PUBLISHER AND AUTHORS HAVE USED THEIR BEST EFFORTS IN PREPARING THIS WORK, THEY MAKE NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE ACCURACY OR COMPLETENESS OF THE CONTENTS OF THIS WORK AND SPECIFICALLY DISCLAIM ALL WARRANTIES, INCLUDING WITHOUT LIMITATION ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. NO WARRANTY MAY BE CREATED OR EXTENDED BY SALES REPRESENTATIVES, WRITTEN SALES MATERIALS OR PROMOTIONAL STATEMENTS FOR THIS WORK. THE FACT THAT AN ORGANIZATION, WEBSITE, OR PRODUCT IS REFERRED TO IN THIS WORK AS A CITATION AND/OR POTENTIAL SOURCE OF FURTHER INFORMATION DOES NOT MEAN THAT THE PUBLISHER AND AUTHORS ENDORSE THE INFORMATION OR SERVICES THE ORGANIZATION, WEBSITE, OR PRODUCT MAY PROVIDE OR RECOMMENDATIONS IT MAY MAKE. THIS WORK IS SOLD WITH THE UNDERSTANDING THAT THE PUBLISHER IS NOT ENGAGED IN RENDERING PROFESSIONAL SERVICES. THE ADVICE AND STRATEGIES CONTAINED HEREIN MAY NOT BE SUITABLE FOR YOUR SITUATION. YOU SHOULD CONSULT WITH A SPECIALIST WHERE APPROPRIATE. FURTHER, READERS SHOULD BE AWARE THAT WEBSITES LISTED IN THIS WORK MAY HAVE CHANGED OR DISAPPEARED BETWEEN WHEN THIS WORK WAS WRITTEN AND WHEN IT IS READ. NEITHER THE PUBLISHER NOR AUTHORS SHALL BE LIABLE FOR ANY LOSS OF PROFIT OR ANY OTHER COMMERCIAL DAMAGES, INCLUDING BUT NOT LIMITED TO SPECIAL, INCIDENTAL, CONSEQUENTIAL, OR OTHER DAMAGES.

For general information on our other products and services, please contact our Customer Care Department within the U.S. at 877-762-2974, outside the U.S. at 317-572-3993, or fax 317-572-4002. For technical support, please visit <https://hub.wiley.com/community/support/dummies>.

Wiley publishes in a variety of print and electronic formats and by print-on-demand. Some material included with standard print versions of this book may not be included in e-books or in print-on-demand. If this book refers to media such as a CD or DVD that is not included in the version you purchased, you may download this material at <http://booksupport.wiley.com>. For more information about Wiley products, visit [www.wiley.com](http://www.wiley.com).

Library of Congress Control Number: 2022950749

ISBN: 978-1-119-98369-9 (pbk); 978-1-119-98370-5 (ebk); 978-1-119-98371-2 (ebk)

# Contents at a Glance

<b>Introduction</b>	1
<b>Book 1: Introducing R</b>	5
CHAPTER 1: R: What It Does and How It Does It	7
CHAPTER 2: Working with Packages, Importing, and Exporting	37
<b>Book 2: Describing Data</b>	51
CHAPTER 1: Getting Graphic	53
CHAPTER 2: Finding Your Center	93
CHAPTER 3: Deviating from the Average	103
CHAPTER 4: Meeting Standards and Standings	113
CHAPTER 5: Summarizing It All	125
CHAPTER 6: What's Normal?	145
<b>Book 3: Analyzing Data</b>	163
CHAPTER 1: The Confidence Game: Estimation	165
CHAPTER 2: One-Sample Hypothesis Testing	181
CHAPTER 3: Two-Sample Hypothesis Testing	207
CHAPTER 4: Testing More than Two Samples	233
CHAPTER 5: More Complicated Testing	257
CHAPTER 6: Regression: Linear, Multiple, and the General Linear Model	279
CHAPTER 7: Correlation: The Rise and Fall of Relationships	315
CHAPTER 8: Curvilinear Regression: When Relationships Get Complicated	335
CHAPTER 9: In Due Time	359
CHAPTER 10: Non-Parametric Statistics	371
CHAPTER 11: Introducing Probability	393
CHAPTER 12: Probability Meets Regression: Logistic Regression	415
<b>Book 4: Learning from Data</b>	423
CHAPTER 1: Tools and Data for Machine Learning Projects	425
CHAPTER 2: Decisions, Decisions, Decisions	449
CHAPTER 3: Into the Forest, Randomly	467
CHAPTER 4: Support Your Local Vector	483
CHAPTER 5: K-Means Clustering	503
CHAPTER 6: Neural Networks	519
CHAPTER 7: Exploring Marketing	537
CHAPTER 8: From the City That Never Sleeps	557

<b>Book 5: Harnessing R: Some Projects to Keep You Busy</b> .....	573
CHAPTER 1: Working with a Browser .....	575
CHAPTER 2: Dashboards — How Dashing! .....	603
<b>Index</b> .....	639

# Table of Contents

<b>INTRODUCTION</b>	1
About This All-in-One	1
Book 1: Introducing R	1
Book 2: Describing Data	2
Book 3: Analyzing Data	2
Book 4: Learning from Data	2
Book 5: Harnessing R: Some Projects to Keep You Busy	3
What You Can Safely Skip	3
Foolish Assumptions	3
Icons Used in This Book	4
Beyond This Book	4
Where to Go from Here	4
 <b>BOOK 1: INTRODUCING R</b>	 5
<b>CHAPTER 1: R: What It Does and How It Does It</b>	7
The Statistical (and Related) Ideas You Just Have to Know	7
Samples and populations	8
Variables: Dependent and independent	8
Types of data	9
A little probability	10
Inferential statistics: Testing hypotheses	12
Null and alternative hypotheses	12
Two types of error	13
Getting R	14
Getting RStudio	15
A Session with R	18
The working directory	18
Getting started	19
R Functions	22
User-Defined Functions	24
Comments	25
R Structures	25
Vectors	25
Numerical vectors	26
Matrices	28
Lists	31
Data frames	32
for Loops and if Statements	35

<b>CHAPTER 2:</b>	<b>Working with Packages, Importing, and Exporting</b>	37
	Installing Packages	37
	Examining Data	39
	Heads and tails	39
	Missing data	40
	Subsets	40
	R Formulas	41
	More Packages	43
	Exploring the tidyverse	44
	Importing and Exporting	47
	Spreadsheets	48
	CSV files	49
	Text files	49
	<b>BOOK 2: DESCRIBING DATA</b>	51
<b>CHAPTER 1:</b>	<b>Getting Graphic</b>	53
	Finding Patterns	53
	Graphing a distribution	54
	Bar-hopping	55
	Slicing the pie	56
	The plot of scatter	57
	Of boxes and whiskers	58
	Doing the Basics: Base R Graphics, That Is	59
	Histograms	60
	Graph features	61
	Bar plots	63
	Pie graphs	65
	Dot charts	65
	Bar plots revisited	66
	Scatter plots	70
	Box plots	73
	Kicking It Up a Notch to ggplot2	73
	Histograms	74
	Bar plots	76
	Dot charts	78
	Bar plots re-revisited	81
	Scatter plots	84
	Box plots	88
	Putting a Bow On It	91
<b>CHAPTER 2:</b>	<b>Finding Your Center</b>	93
	Means: The Lure of Averages	93
	Calculating the Mean	94



	The Average in R: <code>mean()</code> . . . . .	95
	What's your condition? . . . . .	95
	Eliminate \$ signs forthwith( ) . . . . .	96
	Explore the data. . . . .	97
	Outliers: The flaw of averages . . . . .	98
	Medians: Caught in the Middle . . . . .	99
	The Median in R: <code>median()</code> . . . . .	100
	Statistics à la Mode . . . . .	101
	The Mode in R . . . . .	101
<b>CHAPTER 3:</b>	<b>Deviating from the Average</b> . . . . .	103
	Measuring Variation . . . . .	104
	Averaging squared deviations: Variance and how to calculate it. . . . .	104
	Sample variance. . . . .	107
	Variance in R. . . . .	108
	Back to the Roots: Standard Deviation. . . . .	108
	Population standard deviation . . . . .	109
	Sample standard deviation . . . . .	109
	Standard Deviation in R . . . . .	110
	Conditions, conditions, conditions . . . . .	110
<b>CHAPTER 4:</b>	<b>Meeting Standards and Standings</b> . . . . .	113
	Catching Some Zs . . . . .	114
	Characteristics of z-scores . . . . .	114
	Bonds versus the Bambino . . . . .	115
	Exam scores . . . . .	116
	Standard Scores in R. . . . .	116
	Where Do You Stand? . . . . .	119
	Ranking in R . . . . .	119
	Tied scores . . . . .	119
	Nth smallest, Nth largest . . . . .	120
	Percentiles . . . . .	120
	Percent ranks . . . . .	122
	Summarizing . . . . .	123
<b>CHAPTER 5:</b>	<b>Summarizing It All</b> . . . . .	125
	How Many? . . . . .	125
	The High and the Low . . . . .	127
	Living in the Moments . . . . .	127
	A teachable moment. . . . .	128
	Back to descriptives. . . . .	129
	Skewness . . . . .	129
	Kurtosis . . . . .	131

Tuning in the Frequency.....	133
Nominal variables: table() et al.....	134
Numerical variables: hist().....	134
Numerical variables: stem().....	140
Summarizing a Data Frame.....	142
<b>CHAPTER 6: What's Normal?</b> .....	145
Hitting the Curve.....	145
Digging deeper.....	146
Parameters of a normal distribution.....	147
Working with Normal Distributions.....	148
Distributions in R.....	149
Normal density function.....	149
Cumulative density function.....	153
Quantiles of normal distributions.....	156
Random sampling.....	158
Meeting a Distinguished Member of the Family.....	160
The standard normal distribution in R.....	161
Plotting the standard normal distribution.....	162
<b>BOOK 3: ANALYZING DATA</b> .....	163
<b>CHAPTER 1: The Confidence Game: Estimation</b> .....	165
Understanding Sampling Distributions.....	166
An EXTREMELY Important Idea: The Central Limit Theorem.....	167
(Approximately) simulating the central limit theorem.....	168
Predictions of the central limit theorem.....	173
Confidence: It Has Its Limits!.....	175
Finding confidence limits for a mean.....	175
Using R to find the confidence limits for a mean.....	177
Fit to a t.....	178
<b>CHAPTER 2: One-Sample Hypothesis Testing</b> .....	181
Hypotheses, Tests, and Errors.....	181
Hypothesis Tests and Sampling Distributions.....	183
Catching Some Z's Again.....	185
Z Testing in R.....	188
t for One.....	189
t Testing in R.....	190
Working with t-Distributions.....	191
Visualizing t-Distributions.....	192
Plotting t in base R graphics.....	193
Plotting t in ggplot2.....	194
One more thing about ggplot2.....	199

Testing a Variance . . . . .	200
Manufacturing an Example . . . . .	200
Testing in R . . . . .	201
Working with Chi-Square Distributions . . . . .	202
Visualizing Chi-Square Distributions. . . . .	203
Plotting chi-square in base R graphics . . . . .	203
Plotting chi-square in ggplot2 . . . . .	205
<b>CHAPTER 3: Two-Sample Hypothesis Testing . . . . .</b>	<b>207</b>
Hypotheses Built for Two . . . . .	207
Sampling Distributions Revisited . . . . .	208
Applying the central limit theorem . . . . .	209
Zs once more . . . . .	211
Z-testing for two samples in R . . . . .	212
<i>t</i> for Two . . . . .	214
Like Peas in a Pod: Equal Variances . . . . .	214
<i>t</i> -Testing in R . . . . .	216
Working with two vectors. . . . .	216
Working with a data frame and a formula. . . . .	216
Visualizing the results . . . . .	218
Like ps and qs: Unequal variances . . . . .	221
A Matched Set: Hypothesis Testing for Paired Samples . . . . .	222
Paired Sample <i>t</i> -testing in R . . . . .	224
Testing Two Variances . . . . .	224
<i>F</i> testing in R . . . . .	226
<i>F</i> in conjunction with <i>t</i> . . . . .	227
Working with <i>F</i> Distributions . . . . .	227
Visualizing <i>F</i> Distributions . . . . .	228
<b>CHAPTER 4: Testing More than Two Samples . . . . .</b>	<b>233</b>
Testing More than Two . . . . .	233
A thorny problem . . . . .	234
A solution . . . . .	235
Meaningful relationships . . . . .	239
ANOVA in R . . . . .	240
Plotting a boxplot to visualize the data . . . . .	241
After the ANOVA . . . . .	242
Contrasts in R . . . . .	244
Unplanned comparisons . . . . .	245
Another Kind of Hypothesis, Another Kind of Test. . . . .	247
Working with repeated measures ANOVA. . . . .	247
Repeated measures ANOVA in R . . . . .	249
Visualizing the results . . . . .	251
Getting Trendy . . . . .	252
Trend Analysis in R . . . . .	256

<b>CHAPTER 5:</b>	<b>More Complicated Testing</b>	257
	Cracking the Combinations	257
	Interactions	259
	The analysis	259
	Two-Way ANOVA in R	261
	Visualizing the two-way results	263
	Two Kinds of Variables . . . at Once	265
	Mixed ANOVA in R	268
	Visualizing the mixed ANOVA results	270
	After the Analysis	271
	Multivariate Analysis of Variance	272
	MANOVA in R	273
	Visualizing the MANOVA results	275
	After the MANOVA	277
<b>CHAPTER 6:</b>	<b>Regression: Linear, Multiple, and the General Linear Model</b>	279
	The Plot of Scatter	280
	Graphing Lines	281
	Regression: What a Line!	283
	Using regression for forecasting	285
	Variation around the regression line	285
	Testing hypotheses about regression	287
	Linear Regression in R	292
	Features of the linear model	294
	Making predictions	294
	Visualizing the scatterplot and regression line	295
	Plotting the residuals	295
	Juggling Many Relationships at Once: Multiple Regression	297
	Multiple regression in R	299
	Making predictions	300
	Visualizing the 3d scatterplot and regression plane	300
	ANOVA: Another Look	303
	Analysis of Covariance: The Final Component of the GLM	307
	But Wait — There's More	313
<b>CHAPTER 7:</b>	<b>Correlation: The Rise and Fall of Relationships</b>	315
	Understanding Correlation	315
	Correlation and Regression	318
	Testing Hypotheses about Correlation	321
	Is a correlation coefficient greater than zero?	321
	Do two correlation coefficients differ?	322

Correlation in $R$ .....	324
Calculating a correlation coefficient .....	324
Testing a correlation coefficient .....	324
Testing the difference between two correlation coefficients .....	325
Calculating a correlation matrix .....	326
Visualizing correlation matrices .....	326
Multiple Correlation .....	328
Multiple correlation in $R$ .....	329
Adjusting $R$ -squared .....	330
Partial Correlation .....	331
Partial Correlation in $R$ .....	<b>332</b>
Semipartial Correlation .....	333
Semipartial Correlation in $R$ .....	<b>333</b>
 <b>CHAPTER 8: Curvilinear Regression: When Relationships Get Complicated</b> .....	 335
What Is a Logarithm? .....	336
What Is $e$ ? .....	338
Power Regression .....	341
Exponential Regression .....	346
Logarithmic Regression .....	351
Polynomial Regression: A Higher Power .....	354
Which Model Should You Use? .....	357
 <b>CHAPTER 9: In Due Time</b> .....	 359
A Time Series and Its Components .....	359
Forecasting: A Moving Experience .....	363
Forecasting: Another Way .....	366
Working with Real Data .....	368
 <b>CHAPTER 10: Non-Parametric Statistics</b> .....	 371
Independent Samples .....	372
Two samples: Wilcoxon rank-sum test. ....	372
More than two samples: Kruskal-Wallis One-Way ANOVA .....	376
Matched Samples .....	378
Two samples: Wilcoxon matched-pairs signed ranks .....	379
More than two samples: Friedman two-way ANOVA .....	380
More than two samples: Cochran's $Q$ .....	383
Correlation: Spearman's $r_s$ .....	386
Correlation: Kendall's Tau .....	388
A Heads-Up. ....	391

<b>CHAPTER 11: Introducing Probability</b>	393
What Is Probability?	393
Experiments, trials, events, and sample spaces	394
Sample spaces and probability	394
Compound Events	395
Union and intersection	395
Intersection, again	396
Conditional Probability	397
Working with the probabilities	398
The foundation of hypothesis testing	398
Large Sample Spaces	398
Permutations	399
Combinations	400
R Functions for Counting Rules	401
Random Variables: Discrete and Continuous	403
Probability Distributions and Density Functions	403
The Binomial Distribution	406
The Binomial and Negative Binomial in R	407
Binomial distribution	407
Negative binomial distribution	409
Hypothesis Testing with the Binomial Distribution	410
More on Hypothesis Testing: R versus Tradition	412
 <b>CHAPTER 12: Probability Meets Regression: Logistic Regression</b>	415
Getting the Data	418
Doing the Analysis	418
Visualizing the Results	421
 <b>BOOK 4: LEARNING FROM DATA</b>	423
 <b>CHAPTER 1: Tools and Data for Machine Learning Projects</b>	425
The UCI (University of California-Irvine) ML Repository	426
Working with a UCI dataset	426
Cleaning up the data	429
Exploring the data	431
Exploring relationships in the data	432
Introducing the Rattle package	438
Using Rattle with iris	442
Getting and (further) exploring the data	442
Finding clusters in the data	445

<b>CHAPTER 2: Decisions, Decisions, Decisions</b>	449
Decision Tree Components	449
Roots and leaves	450
Tree construction	451
Decision Trees in R	451
Growing the tree in R	452
Drawing the tree in R	453
Decision Trees in Rattle	455
Creating the tree	456
Drawing the tree	457
Evaluating the tree	458
Project: A More Complex Decision Tree	459
The data: Car evaluation	459
Data exploration	461
Building and drawing the tree	462
Evaluating the tree	463
Quick suggested project: Understanding the complexity parameter	464
Suggested Project: Titanic	465
<b>CHAPTER 3: Into the Forest, Randomly</b>	467
Growing a Random Forest	467
Random Forests in R	469
Building the forest	469
Evaluating the forest	470
A closer look	471
Plotting error	473
Plotting importance	475
Project: Identifying Glass	476
The data	476
Getting the data into Rattle	477
Exploring the data	478
Growing the random forest	480
Visualizing the results	480
Suggested Project: Identifying Mushrooms	482
<b>CHAPTER 4: Support Your Local Vector</b>	483
Some Data to Work With	483
Using a subset	484
Defining a boundary	484
Understanding support vectors	485
Separability: It's Usually Nonlinear	486
Support Vector Machines in R	489
Working with e1071	489
Working with kernlab	494

Project: House Parties . . . . .	496
Reading in the data . . . . .	497
Exploring the data . . . . .	499
Creating the SVM . . . . .	500
Evaluating the SVM . . . . .	502
<b>CHAPTER 5: K-Means Clustering . . . . .</b>	<b>503</b>
How It Works . . . . .	503
K-Means Clustering in R . . . . .	505
Setting up and analyzing the data . . . . .	505
Understanding the output . . . . .	506
Visualizing the clusters . . . . .	508
Finding the optimum number of clusters . . . . .	508
Quick suggested project: Adding the sepals . . . . .	513
Project: Glass Clusters . . . . .	514
The data . . . . .	514
Starting Rattle and exploring the data . . . . .	515
Preparing to cluster . . . . .	516
Doing the clustering . . . . .	516
Going beyond Rattle . . . . .	517
<b>CHAPTER 6: Neural Networks . . . . .</b>	<b>519</b>
Networks in the Nervous System . . . . .	519
Artificial Neural Networks . . . . .	520
Overview . . . . .	520
Input layer and hidden layer . . . . .	521
Output layer . . . . .	522
How it all works . . . . .	523
Neural Networks in R . . . . .	523
Building a neural network for the iris data frame . . . . .	523
Plotting the network . . . . .	525
Evaluating the network . . . . .	526
Quick suggested project: Those sepals . . . . .	527
Project: Banknotes . . . . .	527
The data . . . . .	527
Taking a quick look ahead . . . . .	528
Setting up Rattle . . . . .	529
Evaluating the network . . . . .	531
Going beyond Rattle: Visualizing the network . . . . .	531
Suggested Projects: Rattling Around . . . . .	533
<b>CHAPTER 7: Exploring Marketing . . . . .</b>	<b>537</b>
Analyzing Retail Data . . . . .	537
The data . . . . .	538
RFM in R . . . . .	539



Enter Machine Learning . . . . .	546
Working with k-means clustering . . . . .	547
Working with Rattle . . . . .	548
Digging into the clusters . . . . .	550
The clusters and the classes . . . . .	552
Quick suggested project . . . . .	553
Suggested Project: Another Data Set . . . . .	553
<b>CHAPTER 8: From the City That Never Sleeps . . . . .</b>	<b>557</b>
Examining the Data Set . . . . .	557
Warming Up . . . . .	558
Glimpsing and viewing . . . . .	558
Piping, filtering, and grouping . . . . .	559
Visualizing . . . . .	561
Joining . . . . .	562
Quick Suggested Project: Airline Names . . . . .	565
Suggested Project: Departure Delays . . . . .	565
Adding a variable: weekday . . . . .	565
Quick Suggested Project: Analyze Weekday Differences . . . . .	566
Delay, weekday, and airport . . . . .	566
Delay and flight duration . . . . .	570
Suggested Project: Delay and Weather . . . . .	572
 <b>BOOK 5: HARNESSING R: SOME PROJECTS TO KEEP YOU BUSY . . . . .</b>	 <b>573</b>
<b>CHAPTER 1: Working with a Browser . . . . .</b>	<b>575</b>
Getting Your Shine On . . . . .	575
Creating Your First shiny Project . . . . .	576
The user interface . . . . .	579
The server . . . . .	580
Final steps . . . . .	581
Getting reactive . . . . .	582
Working with ggplot . . . . .	585
Changing the server . . . . .	586
A few more changes . . . . .	588
Getting reactive with ggplot . . . . .	590
Another shiny Project . . . . .	592
The base R version . . . . .	593
The ggplot version . . . . .	600
Suggested Project . . . . .	602

<b>CHAPTER 2: Dashboards — How Dashing!</b>	<b>603</b>
The shinydashboard Package	603
Exploring Dashboard Layouts	604
Getting started with the user interface	605
Building the user interface: Boxes, boxes, boxes	605
Lining up in columns	613
A nice trick: Keeping tabs	616
Suggested project: Add statistics	620
Suggested project: Place valueBoxes in tabPanels	621
Working with the Sidebar	622
The user interface	624
The server	626
Suggested project: Relocate the slider	629
Interacting with Graphics	630
Clicks, double-clicks, and brushes — oh, my!	630
Why bother with all this?	634
Suggested project: Experiment with airquality	636
<b>INDEX</b>	<b>639</b>

# Introduction

In this book, I've brought together all the information you need to hit the ground running with R. It's heavy on statistics, of course, because R's creators built this language to analyze data.

So it's necessary that you learn the foundations of statistics. Let me tell you at the outset: This *All-in-One* is not a cookbook. I've never taught statistics that way and I never will. Before I show you how to use R to work with a statistical concept, I give you a strong grounding in what that concept is all about.

In fact, Books 2 and 3 of this 5-book compendium are something like an introductory statistics text that happens to use R as a way of explaining statistical ideas.

Book 4 follows that path by teaching the ideas behind machine learning before you learn how to use R to implement them. Book 5 gives you a set of projects that give you a chance to exercise your newly minted R skill set.

Want some more details? Read on.

## About This All-in-One

The volume you're holding (or the e-book you're viewing) consists of five books that cover a lot of the length and breadth of R.

### Book 1: Introducing R

As I said earlier in this introduction, R is a language that deals with statistics. Accordingly, Book 1 introduces you to the fundamental concepts of statistics that you just *have* to know in order to progress with R.

You then learn about R and RStudio, a widely used development environment for working with R. I begin by describing the rudiments of R code, and I discuss R functions and structures.

R truly comes alive when you use its specialized packages, which you learn about early on.

## **Book 2: Describing Data**

Part of working with statistics is to summarize data in meaningful ways. In Book 2, you find out how to do just that.

Most people know about averages and how to compute them. But that's not the whole story. In Book 2, I tell you about additional descriptive statistics that fill in the gaps, and I show you how to use R to calculate and work with those statistics. You also learn to create graphics that visualize the data descriptions and analyses you encounter in Books 2 and 3.

## **Book 3: Analyzing Data**

Book 3 addresses the fundamental aim of statistical analysis: to go beyond the data and help you make decisions. Usually, the data are measurements of a sample taken from a large population. The goal is to use these data to figure out what's going on in the population.

This opens a wide range of questions: What does an average mean? What does the difference between two averages mean? Are two things associated? These are only a few of the questions I address in Book 3, and you learn to use the R tools that help you answer them.

## **Book 4: Learning from Data**

Effective machine learning model creation comes with experience. Accordingly, in Book 4 you gain experience by completing machine learning projects. In addition to the projects you complete along with me, I suggest additional projects for you to try on your own.

I begin by telling you about the University of California–Irvine Machine Learning Repository, which provides the data sets for most of the projects you encounter in Book 4.

To give you a gentle on-ramp into the field, I show you the `Rattle` package for creating machine learning applications. It's a friendly interface to R's machine learning functionality. I like `Rattle` a lot, and I think you will, too. You use it to learn about and work with decision trees, random forests, support vector machines, k-means clustering, and neural networks.

You also work with fairly large data sets — not the terabytes and petabytes data scientists work with, but large enough to get you started. In one project, you analyze a data set of more than 500,000 airline flights. In another, you complete a customer segmentation analysis of over 300,000 customers of an online retailer.

## Book 5: Harnessing R: Some Projects to Keep You Busy

As its title suggests, Book 5 is also organized around projects.

In these projects, you create applications that respond to users. I show you the `shiny` package for working with web browsers and the `shinydashboard` package for creating dashboards.

All this is a little far afield from R's original mission in life, but you get an idea of R's potential to expand in new directions.

After you've worked with R for a while, maybe you can discover some of those new directions!

## What You Can Safely Skip

Any reference book throws a lot of information at you, and this one is no exception. I intended it all to be useful, but I didn't aim it all at the same level. So if you're not deeply into the subject matter, you can avoid paragraphs marked with the Technical Stuff icon, and you can also skip the sidebars.

## Foolish Assumptions

I'm assuming that you

- » Know how to work with Windows or the Mac. I don't go through the details of pointing, clicking, selecting, and so forth.
- » Can install R and RStudio (I show you how in Book 1) and follow along with the examples. I use the Windows version of RStudio, but you should have no problem if you're working on a Mac.

# Icons Used in This Book

As is the case in all *For Dummies* books, icons help guide you through your journey. Each one is a little picture in the margin that lets you know something special about the paragraph it's next to.



TIP

This icon points out a hint or a shortcut that helps you in your work and makes you an all-around better person.



REMEMBER

This one points out timeless wisdom to take with you as you continue on the path to enlightenment.



WARNING

Pay attention to this icon. It's a reminder to avoid something that might gum up the works for you.



TECHNICAL  
STUFF

As I mention in “What You Can Safely Skip,” this icon indicates material you can blow past if it's just too technical. (I've kept this content to a minimum.)

## Beyond This Book

In addition to what you're reading right now, this book comes with a free, access-anywhere Cheat Sheet that will help you quickly use the tools I discuss. To find this Cheat Sheet, visit [www.dummies.com](http://www.dummies.com) and search for *R All-in-One For Dummies Cheat Sheet* in the Search box.

If you've read any of my earlier books, welcome back!

## Where to Go from Here

Time to hit the books! You can start from anywhere, but here are a couple of hints. Want to introduce yourself to R and packages? Book 1 is for you. Has it been a while (or maybe never?) since your last statistics course? Hit Book 2. For anything else, find it in the table of contents or in the index and go for it.

If you prefer to read from cover to cover, just turn the page. . . .

# 1

## Introducing R

# Contents at a Glance

---

CHAPTER 1:	<b>R: What It Does and How It Does It.....</b>	<b>7</b>
CHAPTER 2:	<b>Working with Packages, Importing, and Exporting .....</b>	<b>37</b>



- » Introducing statistics
- » Getting R and RStudio on your computer
- » Starting a session with R
- » Working with R functions
- » Working with R structures

## Chapter **1**

# R: What It Does and How It Does It

**S**o you're ready to journey into the wonderful world of R! Designed by and for statisticians and data scientists, R has a short but illustrious history.

In the 1990s, Ross Ihaka and Robert Gentleman developed R at the University of Auckland, New Zealand. The R Core Team and the R Foundation for Statistical Computing support R, which has a huge worldwide user base.

Before I tell you about R, however, I have to introduce you to the world that R lives in — the world of data and statistics.

## The Statistical (and Related) Ideas You Just Have to Know

The analytical tools that R provides are based on statistical concepts I help you explore in this section. As you'll see, these concepts are based on common sense.

# Samples and populations

If you watch TV on election night, you know that one of the main events is the prediction of the outcome immediately after the polls close (and before all the votes are counted). How is it that pundits almost always get it right?

The idea is to talk to a *sample* of voters right after they vote. If they're truthful about how they marked their ballots, and if the sample is representative of the *population* of voters, analysts can use the sample data to draw conclusions about the population.

That, in a nutshell, is what statistics is all about — using the data from samples to draw conclusions about populations.

Here's another example. Imagine that your job is to find the average height of 10-year-old children in the United States. Because you probably wouldn't have the time or the resources to measure every child, you'd measure the heights of a representative sample. Then you'd average those heights and use that average as the estimate of the population average.

Estimating the population average is one kind of *inference* that statisticians make from sample data. I discuss inference in more detail in the later section "Inferential Statistics: Testing Hypotheses."



REMEMBER

Here's some important terminology: Properties of a population (like the population average) are called *parameters*, and properties of a sample (like the sample average) are called *statistics*. If your only concern is the sample properties (like the heights of the children in your sample), the statistics you calculate are *descriptive*. (I discuss descriptive statistics in Book 2.) If you're concerned about estimating the population properties, your statistics are *inferential*. (I discuss inferential statistics in Book 3.)



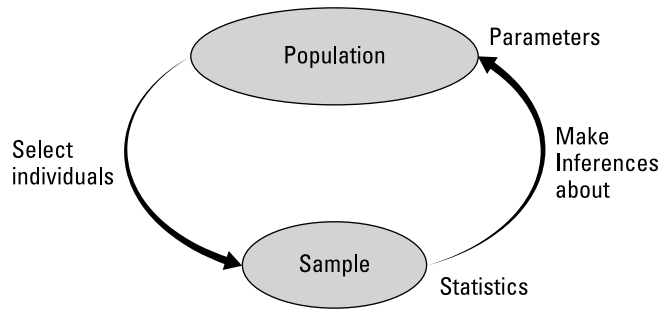
REMEMBER

Now for an important convention about notation: Statisticians use Greek letters ( $\mu$ ,  $\sigma$ ,  $\rho$ ) to stand for parameters, and English letters ( $\bar{X}$ ,  $s$ ,  $r$ ) to stand for statistics. Figure 1-1 summarizes the relationship between populations and samples, and between parameters and statistics.

## Variables: Dependent and independent

A *variable* is something that can take on more than one value — like your age, the value of the dollar against other currencies, or the number of games your favorite sports team wins. Something that can have only one value is a *constant*. Scientists tell us that the speed of light is a constant, and we use the constant  $\pi$  to calculate the area of a circle.

**FIGURE 1-1:**  
The relationship  
between  
populations,  
samples,  
parameters, and  
statistics.



Statisticians work with *independent* variables and *dependent* variables. In any study or experiment, you'll find both kinds. Statisticians assess the relationship between them.

For example, imagine a computerized training method designed to increase a person's IQ. How would a researcher find out whether this method does what it's supposed to do? First, the researcher would randomly assign a sample of people to one of two groups. One group would receive the training method, and the other would complete another kind of computer-based activity — like reading text on a website. Before and after each group completes its activities, the researcher measures each person's IQ. What happens next? I discuss that topic in the later section "Inferential Statistics: Testing Hypotheses."

For now, understand that the independent variable here is Type of Activity. The two possible values of this variable are IQ Training and Reading Text. The dependent variable is the change in IQ from Before to After.



REMEMBER

A dependent variable is what a researcher *measures*. In an experiment, an independent variable is what a researcher *manipulates*. In other contexts, a researcher can't manipulate an independent variable. Instead, they note naturally occurring values of the independent variable and how they affect a dependent variable.



REMEMBER

In general, the objective is to find out whether changes in an independent variable are associated with changes in a dependent variable.

In examples that appear throughout this book, I show you how to use R to calculate characteristics of groups of scores, or to compare groups of scores. Whenever I show you a group of scores, I'm talking about the values of a dependent variable.

## Types of data

When you do statistical work, you can run into four kinds of data. And when you work with a variable, the way you work with it depends on what kind of data it is:

The first kind is *nominal* data. If a set of numbers happens to be nominal data, the numbers are labels — their values don't signify anything. On a sports team, the jersey numbers are nominal. They just identify the players.

The next kind is *ordinal* data. In this data type, the numbers are more than just labels. As the name *ordinal* might tell you, the order of the numbers is important. If I ask you to rank ten foods from the one you like best (1) to the one you like least (10), we'd have a set of ordinal data.

But the difference between your third-favorite food and your fourth-favorite food might not be the same as the difference between your ninth-favorite and your tenth-favorite. So this type of data lacks equal intervals and equal differences.

*Interval* data gives us equal differences. The Fahrenheit scale of temperature is a good example. The difference between 30° and 40° is the same as the difference between 90° and 100°. So each degree is an interval.

People are sometimes surprised to find out that on the Fahrenheit scale a temperature of 80° is not twice as hot as 40°. For ratio statements (“twice as much as,” “half as much as”) to make sense, *zero* has to mean the complete absence of the thing you're measuring. A temperature of 0° F doesn't mean the complete absence of heat — it's just an arbitrary point on the Fahrenheit scale. (The same holds true for Celsius.)

The fourth kind of data, *ratio*, provides a meaningful zero point. On the Kelvin scale of temperature, *zero* means absolute zero, where all molecular motion (the basis of heat) stops. So 200° Kelvin is twice as hot as 100° Kelvin. Another example is length. Eight inches is twice as long as 4 inches. *Zero inches* means a complete absence of length.



REMEMBER

An independent variable or a dependent variable can be either nominal, ordinal, interval, or ratio. The analytical tools you use depend on the type of data you work with.

## A little probability

When statisticians make decisions, they use probability to express their confidence about those decisions. They can never be absolutely certain about what they decide. They can tell you only how probable their conclusions are.

What do we mean by *probability*? Mathematicians and philosophers might give you complex definitions. In my experience, however, the best way to understand probability is in terms of examples.

Here's a simple example: If you toss a coin, what's the probability that it turns up heads? If the coin is fair, you might figure that you have a 50–50 chance of heads and a 50–50 chance of tails. And you'd be right. In terms of the kinds of numbers associated with probability, that's  $\frac{1}{2}$ .

Think about rolling a fair die (one member of a pair of dice). What's the probability that you roll a 4? Well, a die has six faces and one of them is 4, so that's  $\frac{1}{6}$ .

Still another example: Select one card at random from a standard deck of 52 cards. What's the probability that it's a diamond? A deck of cards has four suits, so that's  $\frac{1}{4}$ .

These examples tell you that if you want to know the probability that an event occurs, count how many ways that event can happen and divide by the total number of events that can happen. In the first two examples (heads, 4), the event you're interested in happens only one way. For the coin, we divide 1 by 2. For the die, we divide 1 by 6. In the third example (diamond), the event can happen 13 ways (Ace through King), so we divide 13 by 52 (to get  $\frac{1}{4}$ ).

Now for a slightly more complicated example. Toss a coin and roll a die at the same time. What's the probability of tails and a 4? Think about all the possible events that can happen when you toss a coin and roll a die at the same time. You could have tails and 1 through 6, or heads and 1 through 6. That adds up to 12 possibilities. The tails-and-4 combination can happen only one way. So the probability is  $\frac{1}{12}$ .

In general, the formula for the probability that a particular event occurs is

$$\text{Pr}(\text{event}) = \frac{\text{Number of ways the event can occur}}{\text{Total number of possible events}}$$

At the beginning of this section, I say that statisticians express their confidence about their conclusions in terms of probability, which is why I brought all this up in the first place. This line of thinking leads to *conditional* probability — the probability that an event occurs given that some other event occurs. Suppose that I roll a die, look at it (so that you don't see it), and tell you that I rolled an odd number. What's the probability that I've rolled a 5? Ordinarily, the probability of a 5 is  $\frac{1}{6}$ , but "I rolled an odd number" narrows it down. That piece of information eliminates the three even numbers (2, 4, 6) as possibilities. Only the three odd numbers (1, 3, 5) are possible, so the probability is  $\frac{1}{3}$ .

What's the big deal about conditional probability? What role does it play in statistical analysis? Read on.

# Inferential statistics: Testing hypotheses

Before a statistician does a study, they draw up a tentative explanation — a *hypothesis* that tells why the data might come out a certain way. After gathering all the data, the statistician has to decide whether to reject the hypothesis.

That decision is the answer to a conditional probability question — what’s the probability of obtaining the data, given that this hypothesis is correct? Statisticians have tools that calculate the probability. If the probability turns out to be low, the statistician rejects the hypothesis.

Back to coin-tossing for an example: Imagine that you’re interested in whether a particular coin is fair — whether it has an equal chance of heads or tails on any toss. Let’s start with “The coin is fair” as the hypothesis.

To test the hypothesis, you’d toss the coin a number of times — let’s say 100. These 100 tosses are the sample data. If the coin is fair (as per the hypothesis), you’d expect 50 heads and 50 tails.

If it’s 99 heads and 1 tail, you’d surely reject the fair-coin hypothesis: The conditional probability of 99 heads and 1 tail given a fair coin is very low. Of course, the coin could still be fair and you could, quite by chance, get a 99-1 split, right? Sure. You never really know. You have to gather the sample data (the 100 toss results) and then decide. Your decision might be right, or it might not.

Juries make these types of decisions. In the United States, the starting hypothesis is that the defendant is not guilty (“innocent until proven guilty”). Think of the evidence as data. Jury members consider the evidence and answer a conditional probability question: What’s the probability of the evidence, given that the defendant is not guilty? Their answer determines the verdict.

## Null and alternative hypotheses

Think again about that coin-tossing study I just mentioned. The sample data are the results from the 100 tosses. I said that we can start with the hypothesis that the coin is fair. This starting point is called the *null hypothesis*. The statistical notation for the null hypothesis is  $H_0$ . According to this hypothesis, any heads-tails split in the data is consistent with a fair coin. Think of it as the idea that nothing in the sample data is out of the ordinary.

An alternative hypothesis is possible — that the coin isn’t a fair one and it’s biased to produce an unequal number of heads and tails. This hypothesis says that any heads-tails split is consistent with an unfair coin. This alternative hypothesis is