Probability and Statistics for Data Science

This self-contained guide introduces two pillars of data science, probability theory, and statistics, side by side, in order to illuminate the connections between statistical techniques and the probabilistic concepts they are based on. The topics covered in the book include random variables, nonparametric and parametric models, correlation, estimation of population parameters, hypothesis testing, principal component analysis, and both linear and nonlinear methods for regression and classification. Examples throughout the book draw from real-world datasets to demonstrate concepts in practice and confront readers with fundamental challenges in data science, such as overfitting, the curse of dimensionality, and causal inference. Code in Python reproducing these examples is available on the book's website, along with videos, slides, and solutions to exercises.

This accessible book is ideal for undergraduate and graduate students, data science practitioners, and others interested in the theoretical concepts underlying data science methods.

CARLOS FERNANDEZ-GRANDA is Associate Professor of mathematics and data science at New York University, where he has taught probability and statistics to data science students since 2015. The goal of his research is to design and analyze data science methodology, with a focus on machine learning, artificial intelligence, and their application to medicine, climate science, biology, and other scientific domains.

"Fernandez-Granda's *Probability and Statistics for Data Science* is a comprehensive yet approachable treatment of the fundamentals required of all aspiring data scientists – whether they be in academia, industry, or elsewhere. The language is clear and precise, and it is one of the best-organized treatments of this material I have ever seen. With lucid examples and helpful exercises, it deserves to be the leading text for these topics among undergraduate and graduate students in this technical, fast-moving discipline. Instructors take note!"

—Arthur Spirling, Princeton University

"If you're mathematically inclined and want to master the foundations of data science in one go, this book is for you. It covers a broad range of essential modern topics – including nonparametric methods, causal inference, latent variable models, Bayesian approaches, and a thorough introduction to machine learning – all illustrated with an abundance of figures and real-world data examples. Highly recommended."

—David Rosenberg, Office of the CTO, Bloomberg

Probability and Statistics for Data Science

CARLOS FERNANDEZ-GRANDA
New York University

CAMBRIDGE
UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre,
New Delhi – 110025, India

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment,
a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of
education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781009180085

DOI: 10.1017/9781009180108

When citing this work, please include a reference to the
DOI 10.1017/9781009180108

First published 2025

Cover image: philipp igumnov/Moment via Getty Images

A catalogue record for this publication is available from the British Library

A Cataloging-in-Publication data record for this book is available from the Library of Congress

ISBN 978-1-009-18008-5 Hardback
ISBN 978-1-009-18009-2 Paperback

Para Irida, Iliana y Chrysa

Contents

Preface		*page* xi
Book Website		xiii
	Introduction and Overview	1
1	**Probability**	6
1.1	Intuitive Properties of Probability	6
1.2	Mathematical Definition of Probability	9
1.3	Conditional Probability	13
1.4	Estimating Probabilities from Data	19
1.5	Independence	22
1.6	Conditional Independence	26
1.7	The Monte Carlo Method	29
	Exercises	33
2	**Discrete Variables**	37
2.1	Discrete Random Variables	37
2.2	The Empirical Probability Mass Function	43
2.3	Discrete Parametric Distributions	46
2.4	Maximum-Likelihood Estimation	52
2.5	Comparing Parametric and Nonparametric Models	57
	Exercises	63
3	**Continuous Variables**	66
3.1	Continuous Random Variables	66
3.2	The Cumulative Distribution Function	69
3.3	The Probability Density Function	76
3.4	Functions of Random Variables	81
3.5	Nonparametric Probability-Density Estimation	84
3.6	Continuous Parametric Distributions	90
3.7	Maximum-Likelihood Estimation	96
3.8	Inverse-Transform Sampling	101
	Exercises	104
4	**Multiple Discrete Variables**	109
4.1	Multivariate Discrete Random Variables	109
4.2	Marginal Distributions	115

4.3	Conditional Distributions	118
4.4	Independence	122
4.5	Conditional Independence	125
4.6	Causal Inference	127
4.7	The Curse of Dimensionality	137
4.8	Classification via Naive Bayes	138
4.9	Markov Chains	141
	Exercises	152
5	**Multiple Continuous Variables**	161
5.1	Joint Distribution of Continuous Random Variables	161
5.2	Joint Cumulative Distribution Function	163
5.3	Joint Probability Density Function	166
5.4	Multidimensional Density Estimation	170
5.5	Marginal Distributions	171
5.6	Conditional Distributions	174
5.7	Independence	177
5.8	Conditional Independence	180
5.9	Jointly Simulating Multiple Random Variables	184
5.10	Gaussian Random Vectors	186
	Exercises	196
6	**Discrete and Continuous Variables**	202
6.1	Joint Distribution of Discrete and Continuous Variables	202
6.2	Conditional Distribution of Continuous Variables Given Discrete Variables	204
6.3	Conditional Distribution of Discrete Variables Given Continuous Variables	207
6.4	Independence	211
6.5	Classification via Gaussian Discriminant Analysis	212
6.6	Clustering via Gaussian Mixture Models	217
6.7	Bayesian Parametric Modeling	225
	Exercises	235
7	**Averaging**	241
7.1	The Mean	241
7.2	Linearity of Expectation	248
7.3	Independent Random Variables	249
7.4	Mean of Parametric Distributions	251
7.5	Sensitivity of the Mean to Extreme Values	254
7.6	The Mean Square	255
7.7	The Variance	257
7.8	The Conditional Mean	262
7.9	The Average Treatment Effect	274
	Exercises	280

8	**Correlation**	284
8.1	Correlation between Standardized Quantities	284
8.2	Correlation and Covariance	287
8.3	Estimating Correlation from Data	290
8.4	Simple Linear Regression	293
8.5	Properties of the Correlation Coefficient	299
8.6	Uncorrelation and Independence	306
8.7	A Geometric Analysis of Correlation	310
8.8	Correlation (Usually) Does Not Imply Causation	315
	Exercises	321
9	**Estimation of Population Parameters**	325
9.1	Random Sampling	325
9.2	The Bias	329
9.3	The Standard Error	332
9.4	Deviation Bounds: Markov's and Chebyshev's Inequalities	337
9.5	The Law of Large Numbers	339
9.6	Some Averages Are Not to Be Trusted	344
9.7	The Central Limit Theorem	349
9.8	Confidence Intervals	363
9.9	The Bootstrap	370
	Exercises	384
10	**Hypothesis Testing**	390
10.1	Selecting a Hypothesis	390
10.2	The Null Hypothesis and the Test Statistic	391
10.3	Parametric Testing and the P-Value	392
10.4	Two-Sample Tests	395
10.5	Statistical Significance	399
10.6	The Power	402
10.7	Nonparametric Testing: The Permutation Test	408
10.8	Multiple Testing	417
10.9	Hypothesis Testing and Causal Inference	423
10.10	P-Value Abuse	425
	Exercises	429
11	**Principal Component Analysis and Low-Rank Models**	433
11.1	The Mean in Multiple Dimensions	433
11.2	The Covariance Matrix	436
11.3	The Sample Covariance Matrix	441
11.4	Principal Component Analysis	445
11.5	Dimensionality Reduction	457
11.6	Low-Rank Models	464
11.7	Matrix Completion for Collaborative Filtering	483
	Exercises	491

12 Regression and Classification 495
12.1 Linear Regression 495
12.2 Generalization and Overfitting in Linear Regression 514
12.3 Ridge Regression 531
12.4 Sparse Regression 543
12.5 Logistic Regression 546
12.6 Softmax Regression 552
12.7 Tree-Based Models 556
12.8 Neural Networks and Deep Learning 577
12.9 Evaluation of Classification Models 584
Exercises 590

Appendix: Datasets 599
References 602
Index 604

Preface

I started teaching probability and statistics to data science students in 2015. At first, I followed a traditional approach. I would begin with probability theory and then eventually move on to statistics and data analysis. It did not work. Students found it hard to understand the connection between the two halves of the course. By the time we reached statistics, they would struggle to recall the relevant concepts from probability theory, and I would often have to review them all over again. Motivated by this, I decided to completely restructure the material and write this book.

The book is a self-contained, rigorous guide to probability and statistics for data science. It intertwines the material on probability and statistics, explaining how to estimate each probabilistic object from data *right after its mathematical definition*. For example, we present nonparametric and parametric models just after defining random variables. This makes it possible to provide examples with real data from the very beginning. Throughout the book, these examples illustrate the connection between the material and practical data analysis, confronting students with fundamental challenges in data science, such as overfitting, the curse of dimensionality, and causal inference.

I began to think about how to best teach probability and statistics as a teaching assistant for Ayfer Özgür and Abbas El Gamal in graduate school. Abbas gave me the opportunity to teach my first course and encouraged me to use real-data examples, one of which survives in this book (the call-center example in Chapter 2). The book is also influenced by several other great teachers I had the luck to learn from in my studies, including Julian Manley, Eduardo Lorenzo Pigueiras, Francis Bach, Dan Boneh, Stephen Boyd, and Andrea Montanari.

While writing the book, I received invaluable feedback from my students at the Center for Data Science in New York University[1]. I had the good fortune to co-teach this material with Brett Bernstein and Ilias Zadik and to have the help of my wonderful teaching assistants Annika Brundyn, Nadine Hussami, Vlad Kobzar, Nhung Le, Mu Li, Xintong Li, Sheng Liu, Taro Makino, Sreyas Mohan, Tinatin Nikvashvili, Jonas Peeters, Lisa Ren, Levent Sagun, Michael Stanley, Zhiyuan Wang, and Weicheng Zhu. I am extremely grateful to Diego Herrero Quevedo and Louis Mittel, who proofread the manuscript and provided many insightful comments, and also to Juan Argote, Afonso Bandeira, Chris Chen, Shi-ang Chi, Daniel Climent, Rajesh Ranganath, Alvaro Reig, and David Rosenberg for their helpful suggestions.

In addition, I would like to thank Diana Gillooly, who originally suggested that I publish the book, my editor Lauren Cowles, for her help and extreme patience, my mentors Emmanuel Candès, Julia Kempe, Bob Kohn, and Eero Simoncelli, for their guidance, and

[1] My favorite feedback from a student evaluation: *The instructor is like Spanish wine. I don't like Spanish wine.*

the Division of Mathematical Sciences of the National Science Foundation, for their support through grants 1616340 and 2009752.

Finally, I cannot end without thanking my in-laws Dimitra and Daniil, my parents Yolanda and Carlos, my daughter Irida (for constantly sharing the books she was writing in kindergarten to motivate me), my daughter Iliana (for showing up as I was writing the book, and making things more fun), and my wife Chrysa (for everything).

Book Website

Supporting material for the book, including code, videos, slides, and solutions to all exercises, can be found on the book website:

www.ps4ds.net