Contents

About the Author

Dr. Norm Matloff is a professor of computer science at the University of California at Davis, and was formerly a professor of statistics at that university. He is a former database software developer in Silicon Valley, and has been a statistical consultant for firms such as the Kaiser Permanente Health Plan.

Dr. Matloff was born in Los Angeles, and grew up in East Los Angeles and the San Gabriel Valley. He has a PhD in pure mathematics from UCLA, specializing in probability theory and statistics. He has published numerous papers in computer science and statistics, with current research interests in parallel processing, regression methodology, machine learning and recommender systems.

Professor Matloff is an award-winning expositor. He is a recipient of the campuswide Distinguished Teaching Award at his university, and his book, *Statistical Regression and Classification: From Linear Models to Machine Learning*, was selected for the international 2017 Ziegel Award. (He also has been a recipient of the campuswide Distinguished Public Service Award at UC Davis.)

To the Instructor

Statistics is not a discipline like physics, chemistry or biology where we study a subject to solve problems in the same subject. We study statistics with the main aim of solving problems in other disciplines — C.R. Rao, one of the pioneers of modern statistics

The function of education is to teach one to think intensively and to think critically. Intelligence plus character — that is the goal of true education — Dr. Martin Luther King, American civil rights leader

[In spite of] innumerable twists and turns, the Yellow River flows east — Confucius, ancient Chinese philosopher

This text is designed for a junior/senior/graduate-level based course in probability and statistics, *aimed specifically at data science students (including computer science).* In addition to calculus, the text assumes some knowledge of matrix algebra and rudimentary computer programming.

But why is this book different from all other books on math probability and statistics?

Indeed. it *is* quite different from the others. Briefly:

- The subtitle of this book, *Math + R + Data*, immediately signals a difference from other "math stat" books.

- Data Science applications, e.g. random graph models, power law distribution, Hidden Markov models, PCA, Google PageRank, remote sensing, mixture distributions, neural networks, the Curse of Dimensionality, and so on.

- Extensive use of the R language.

The subtitle of this book, *Math + R + Data*, immediately signals that the book follows a very different path. Unlike other "math stat" books, this one has a strong applied emphasis, with lots of real data, facilitated by extensive use of the R language.

The above quotations explain the difference further. First, this book is definitely written from an applications point of view. Second, it pushes the student to think critically about the *how* and *why* of statistics, and to "see the big picture."

- **Use of real data, and early introduction of statistical issues:**

  The Rao quote at the outset of this Preface resonates strongly with me. Though this is a "math stat" book — random variables, density functions, expected values, distribution families, stat estimation and inference, and so on — it takes seriously the Data Science theme claimed in the title, *Probability and Statistics for Data Science*. A book on Data Science, even a mathematical one, should make heavy use of DATA!

  This has implications for the ordering of the chapters. We bring in statistics early, and statistical issues are interspersed thoughout the text. Even the introduction to expected value, Chapter 3, includes a simple prediction model, serving as a preview of what will come in Chapter 15. Chapter 5, which covers the famous discrete parametric models, includes an example of fitting the power law distribution to real data. This forms a prelude to Chapter 7, which treats sampling distributions, estimation of mean and variance, bias and so on. Then Chapter 8 covers general point estimation, using MLE and the Method of Moments to fit models to real data. From that point onward, real data is used extensively in every chapter.

  The datasets are all publicly available, so that the instructor can delve further into the data examples.

- **Mathematically correct – yet highly intuitive:**

  The Confucius quote, though made long before the development of formal statistical methods, shows that he had a keen **intuition**, anticipating a fundamental concept in today's world of data science — data smoothing. Development of such strong intuition in our students is a high priority of this book.

  This is of course a mathematics book. All models, concepts and so on are described precisely in terms of random variables and distributions. In addition to calculus, matrix algebra plays an important role. Optional Mathematical Complements sections at the ends of many chapters allow inquisitive readers to explore more sophisticated material. The mathematical exercises range from routine to more challenging.

On the other hand, this book is not about "math for math's sake." In spite of being mathematically precise in description, it is definitely not a theory book.

For instance, the book does not define probability in terms of sample spaces and set-theoretic terminology. In my experience, defining probability in the classical manner is a major impediment to learning the intuition underlying the concepts, and later to doing good applied work. Instead, I use the intuitive, informal approach of defining probability in terms of long-run frequency, in essence taking the Strong Law of Large Numbers as an axiom.

I believe this approach is especially helpful when explaining conditional probability and expectation, concepts that students notoriously have trouble with. Under the classical approach, students have trouble recognizing when an exercise — and more importantly, an actual application — calls for a conditional probability or expectation if the wording lacks the explicit phrase *given that*. Instead, I have the reader think in terms of repeated trials, "How often does A occur *among those times* in which B occurs?", which is easier to relate to practical settings.

- **Empowering students for real-world applications:**

  The word *applied* can mean different things to different people. Consider for instance the interesting, elegant book for computer science students by Mitzenmacher and Upfal [33]. It focuses on probability, in fact discrete probability, and its intended class of applications is actually the *theory* of computer science.

  I instead focus on the actual use of the material in the real world; which tends to be more continuous than discrete, and more in the realm of statistics than probability. This is especially valuable, as Big Data and Machine Learning now play a significant role in computer and data science.

  One sees this philosophy in the book immediately. Instead of starting out with examples involving dice or coins, the book's very first examples involve a model of a bus transportation system and a model of a computer network. There are indeed also examples using dice, coins and games, but the theme of the late Leo Breiman's book subtitle [5], "With a View toward Applications," is never far away.

If I may take the liberty of extending King's quote, I would note that today statistics is a core intellectual field, affecting virtually everyone's daily lives. The ability to use, or at the very least *understand*, statistics is vital to good citizenship, and as an author I take this as a mission.

- **Use of the R Programming Language:**

  The book makes use of some light programming in R, for the purposes of simulation and data analysis. The student is expected to have had some rudimentary prior background in programming, say in one of Python, C, Java or R, but no prior experience with R is assumed. A brief introduction is given in the book's appendix, and some further R topics are interspered with the text as Computational Complements.

  R is widely used in the world of statistics and data science, with outstanding graphics/visualization capabilities, and a treasure chest of more than 10,000 contributed code packages.

  Readers who happen to be in computer science will find R to be of independent interest from a CS perspective. First, R follows the *functional language* and *object-oriented* paradigms: Every action is implemented as a function (even '+'); side effects are (almost) always avoided; functions are first-class objects; several different kinds of class structures are offered. R also offers various interesting metaprogramming capabilities. In terms of programming support, there is the extremely popular RStudio IDE, and for the "hard core" coder, the Emacs Speaks Statistics framework. Most chapters in the book have Computational Complements sections, as well as a Computational and Data Problems portion in the exercises.

Chapter Outline:

Part I, Chapters 1 through 6: These introduce probability, Monte Carlo simulation, discrete random variables, expected value and variance, and parametric families of discrete distributions.

Part II, Chapters 7 through 10: These then introduce statistics, such as sampling distributions, MLE, bias, Kolmogorov-Smirnov and so on, illustrated by fitting gamma and beta density models to real data. Histograms are viewed as density estimators, and kernel density estimation is briefly covered. This is followed by material on confidence intervals and significance testing.

Part III, Chapters 11 through 17: These cover multivariate analysis in various aspects, such as multivariate distribution, mixture distributions, PCA/log-linear model, dimension reduction, overfitting and predictive analytics. Again, real data plays a major role.

Coverage Strategies:

The book can be comfortably covered in one semester. If a more leisurely pace is desired, or one is teaching under a quarter system, the material has been designed so that some parts can be skipped without loss of continuity. In particular, a more statistics-oriented course might omit the material on Markov chains, while a course focusing more on machine learning may wish to retain this material (e.g. for Hidden Markov models). Individual sections on specialty topics also have been written so as not to create obstacles later on if they are skipped.

The Chapter 11 on multivariate distributions is very useful for data science, e.g. for its relation to clustering. However, instructors who are short on time or whose classes may not have a strong background in matrix algebra may safely skip much of this material.

A Note on Typography

In order to help the reader keep track of the various named items, I use math italics for mathematical symbols and expressions, and bold face for program variable and function names. I include R package names for the latter, except for those beginning with a capital letter.

Thanks:

The following, among many, provided valuable feedback for which I am very grateful: Ibrahim Ahmed; Ahmed Ahmedin; Stuart Ambler; Earl Barr; Benjamin Beasley; Matthew Butner; Vishal Chakraborti, Michael Clifford; Dipak Ghosal; Noah Gift; Laura Matloff; Nelson Max, Deep Mukhopadhyay, Connie Nguyen, Jack Norman, Richard Oehrle, Michael Rea, Sana Vaziri, Yingkang Xie, and Ivana Zetko. My editor, John Kimmel, is always profoundly helpful. And as always, my books are also inspired tremendously by my wife Gamis and daughter Laura.

To the Reader

I took a course in speed reading, and read War and Peace in 20 minutes. It's about Russia — comedian Woody Allen

I learned very early the difference between knowing the name of something and knowing something — Richard Feynman, Nobel laureate in physics

Give me six hours to chop down a tree and I will spend the first four sharpening the axe — Abraham Lincoln

This is NOT your ordinary math or programming book.

In order to use this material in real-world applications, it's crucial to understand what the math *means*, and what the code actually *does*.

In this book, you will often find several consecutive paragraphs, maybe even a full page, in which there is no math, no code and no graphs. Don't skip over these portions of the book! They may actually be the most important ones in the book, in terms of your ability to apply the material in the real world.

And going hand-in-hand with this point, mathematical intuition is key. As you read, stop and think about the intuition underlying those equations.

A closely related point is that the math and code complement each other. Each will give you deeper insight in the other. It may at first seem odd that the book intersperses math and code, but sooon you will find their interaction to be quite helpful to your understanding of the material.

The "Plot"

Think of this book as a movie. In order for the "plot" to work well, we will need preparation. This book is aimed at applications to Data Science, so the ultimate destination of the "plot" is statistics and predictive analytics.

The foundation for those fields is probability, so we lay the foundation first in Chapters 1 through 6. We'll need more probability later — Chapters 9 and 11 — but in order to bring in some "juicy" material into the "movie" as early as possible, we introduce statistics, especially analysis of real DATA, in Chapters 7 and 8 at this early stage.

The final chapter, on Markov chains, is like a "sequel" to the movie. This sets up some exciting Data Science applications such as Hidden Markov Models and Google's PageRank search engine.