Off the Beaten Path Tutorial: Stochastic Processes and Simulations – Volume 1

Vincent Granville, Ph.D.
Data Shaping Solutions, LLC

Anacortes, WA, February 2022

Note: External links (in blue) and internal references (in red) are clickable throughout this document. Keywords highlighted in orange are indexed; those in red are both indexed and in the glossary section.

Contents

1 About this Textbook

This section answers the following questions: what you will learn from the textbook, the prerequisites, who should buy this book, and why. It also features a short bio of the author.

1.1 Brief Presentation

This scratch course on stochastic processes covers significantly more material than usually found in traditional books or classes. The approach is original: I introduce a new yet intuitive type of random structure called perturbed lattice or Poisson-binomial process, as the gateway to all the stochastic processes. Such models have started to gain considerable momentum recently, especially in sensor data, cellular networks, chemistry, physics

and engineering applications. I present state-of-the-art material in simple words, in a compact style, including new research developments and open problems. I focus on the methodology and principles, providing the reader with solid foundations and numerous resources: theory, applications, illustrations, statistical inference, references, glossary, educational spreadsheet, source code, stochastic simulations, original exercises, videos and more.

Highlights

Below is a short selection of topics covered in the textbook. Some are research results published here for the first time.

- New probability distributions: exponential-binomial with infinitely many parameters, generalized logistic
- Fractal supervised clustering in GPU (graphics processing unit) using image filtering techniques
- Unsupervised clustering in GPU using density (gray levels) equalizer
- Automated, black-box detection of the number of clusters
- Rayleigh test, new test of independence, inference and model fitting techniques for spatial processes
- Consolidation of the three attractor distributions in extreme value theory
- Well documented source code and formulas to generate various deviates and 2D simulations
- Detailed and simple recipes to design your own data animations as mp4 videos
- Complex mixture and surperimposed models, hexagonal lattices, coverage problems
- Properties of nearest neighbor graphs, size distribution of connected components
- Central limit theorem for Poisson-binomial point processes, convergence and numerical stability issues
- Chaotic convergence of perturbed series in experimental number theory, related to the Riemann Hypothesis
- Generation of random functions, random graphs, random permutations, random numbers and more
- 25 Exercises with solution, extending the theory and methods developed in the textbook

This first textbook deals with point processes in one and two dimensions, including spatial processes and clustering. The next book in this series will cover other types of stochastic processes, such as Brownian-related and random, chaotic dynamical systems. The point process which is at the core of this textbook is called the Poisson-binomial process (not to be confused with a binomial nor a Poisson process) for reasons that will soon become apparent to the reader. Two extreme cases are the standard Poisson process, and fixed (non-random) points on a lattice. Everything in between is the most exciting part.

1.2 Target Audience

College-educated professionals with an analytical background (physics, economics, finance, machine learning, statistics, computer science, quant, mathematics, operations research, engineering, business intelligence), students enrolled in a quantitative curriculum, decision makers or managers working with data scientists, graduate students, researchers and college professors, will benefit the most from this textbook. The textbook is also intended to professionals interested in automated machine learning and artificial intelligence.

Most references are accessible online with one click: they are highlighted in blue in the text. Many original exercises requiring out-of-the-box thinking, are offered with solution. Professors can include them in their classes or exams. Students should try them. Difficult ones are starred to warn the reader. Some are rather theoretical, and some involve writing code to perform simulations and test some hypotheses. Most of these exercises are an extension of the core material presented in this textbook.

The textbook includes full source code, in particular for simulations, image processing, and video generation. You don't need to be a programmer to understand the code. It is well documented and easy to read, even for people with little or no programming experience. Indeed, the format is designed to help you quickly develop and implement your own machine learning applications from scratch, whether basic or advanced, with relatively simple and limited code. Emphasis is on good coding practices. The textbook also features professional-looking spreadsheets allowing you to perform interactive statistical tests and simulations in Excel alone, without statistical tables or any coding.

The code, data sets and spreadsheets are available on my GitHub repository. GitHub is the platform of choice for developers. The architecture of my repository is well thought out, and can inspire you if/when you build your own. Without having to spend time and money on classes to get started, this textbook can help you jump-start a well paid developer career.

Finally, by focusing on high value material only, the textbook is short and easy to read for busy professionals. Basic or side material is accessible via clickable links to well-written online references, throughout the textbook. For instance, when mentioning "numerical stability", the reader is referred to the Wikipedia entry for details,

denoted as [Wiki], and accessible from the PDF with one click. Important concepts or topics, discussed in detail in the textbook, are highlighted in dark orange and listed in the index: for instance, stationarity. The reader is invited to check the index, to find relevant cross-references to the concept in question, within the textbook. Internal references to specific sections, figures, tables, theorems, exercises, formulas and so on, are highlighted in red and clickable. Also, the bibliography has back links to the citations in the textbook. In short, the whole design of the PDF makes navigation and searches easy.

1.3 List of Exercises

Section 6 consists of exercises that complement and significantly expand the material discussed in this textbook. Even if you don't have time to solve them, I encourage you to a least browse them. Solutions are provided, and some exercises involve simulations only. Starred exercises are more difficult. Here, NN stands for nearest neighbors.

1	Point count, Laplace distribution	14	Distribution of NN distances
2 *	Convergence to Poisson process	15	Cell networks: coverage problem
3 *	Limit of generalized logistic distribution	16	Optimum circle covering of the plane
4	Small paradox	17	Interlaced lattices, lattice mixtures, NN
5	Exact distribution of interarrival times	18 *	Lattice topology and algebra
6 *	Retrieving F from interarrival times	19 **	NN graph: size of connected components
7 *	Poisson limit of Poisson-binomial distribution	20	NN graph: maximum clique problem
8	A few simple theorems	21	Computing moments using the CDF
9	Ergodicity, independent increments	22	Simulations: generalized logistic distribution
10	Boundary effect	23	Riemann Hypothesis
11	A curious, Poisson-like point process	24 *	Convergence acceleration
12 *	Poisson-binomial process on the sphere	25	Fast image filtering algorithm
13	Taxonomy of point processes		

1.4 About the Author

Vincent Granville, PhD is a pioneering data scientist and machine learning expert, co-founder of Data Science Central (acquired by a publicly traded company in 2020), former VC-funded executive, author and patent owner. Vincent's past corporate experience includes Visa, Wells Fargo, eBay, NBC, Microsoft, CNET, InfoSpace and other Internet startup companies (one acquired by Google). Vincent is also a former post-doct from Cambridge University, and the National Institute of Statistical Sciences (NISS). He is currently publisher at DataShaping.com. He makes a living as an independent researcher working on stochastic processes, dynamical systems, experimental math and probabilistic number theory.

Vincent published in Journal of Number Theory, Journal of the Royal Statistical Society (Series B), and IEEE Transactions on Pattern Analysis and Machine Intelligence, among others. He is also the author of multiple books, including "Statistics: New Foundations, Toolbox, and Machine Learning Recipes", "Applied Stochastic Processes, Chaos Modeling, and Probabilistic Properties of Numeration Systems" with a combined reach of over 250,000, as well as "Becoming a Data Scientist" published by Wiley. For details, see my Google Scholar profile, here.