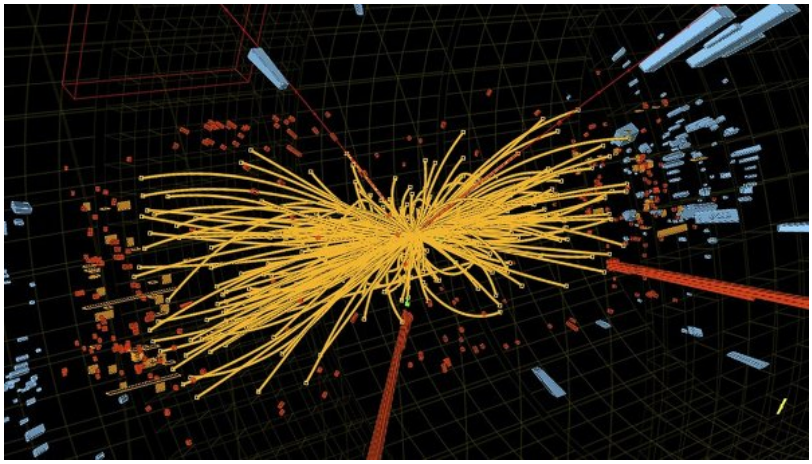
Alessandro Candolini

Statistical Data Analysis

Theory, methods, algorithms and modern techniques
from high-energy and computational physics to information technology



March 14, 2021

E-mail: alessandro.candolini@gmail.com

First printing, September 2014

“There is nothing more practical than a good theory.”

— Kurt Lewin, 1945

CONTENTS

PREFACE

This preface and these notes are written from a physicist' perspective. This will ultimately drive the approach towards the subject and it will bias the selection of topics presented here, the examples discussed throughout this work, and the level of mathematical rigour (higher than what is customary in several books of practical statistics) , leading to prioritize definitely what is likely to be of major relevance or more common for an audiance of physicists, nevertheless trying to emphasize at the same time the impacts, implications and applications in contemporary information technology and big data community.

A through understanding of traditional and modern statistical inference (theory, methodologies, techniques, and algorithms) has always been playing quite a prominent role in the education and training of a scientist (in general) and of a physicist (in particular), providing the framework and the tools to analyse in a reliable and rigorous way the outcomes of either actual experiments or virtual computer simulations, in order to get insights, estimate parameters and their accuracy and precision, make comparisons between theory and experiments, validate hypothesis tests, assess the validity of a theoretical model and its underlying assumptions against experimental data, and make predictions.

The need of a powerful background in statistics has probably become even more demanding in recent years for practitioners in data analysis. The exponential growth of computing resources during the past few decades has made 21^{st} century increasingly data-intensive. Beyond "traditional methods (e. g., probabilitic modelling, linear and non-linear regression theory, confidence intervals, theory of estimators, statistical significance, hypothesis test, time series analysis, etc), new computer-intensive powerful techniques has emerged in the last decades, including new progress in Bayesian statistics, maximum entropy methods, computational statistics (montecarlo methods, expectation-maximization algorithm, hidden Markov models, resampling techniques such as the statistical bootstrap, etc), statistical learning (artificial neural networks, support vector machines, k-means clustering, and related topics in pattern recognition, etc), just to mention few examples.

In this respect, physics has always been a rich source of probabilistic models. (statistical mechanics, etc). statistical mechanics (lattice spin models such as Potts models or spin glasses, ergodic theory), group theoretical methods and concepts borrower from relativistic quantum field theory and high energy particle physics provide valuable insights to better understand the mechanism behind these tools.

From a theoretical viewpoint, is a beautiful and fashinating field of research, sharing important connections with information theory, mathematics, artificial intelligence and robotics, physics (e. g., statistical mechanics), etc. But the relevance of these methods is far from being just of theoretical interest.

From a practical viewpoint, statistical methods are undergoing a tremendus development in recent years. This is not only relevant for physics however. playing an increasing widespread role in modern information technology and big data community, Warning about the misleading analysis and abuse of some statistical technique, nevertheless helping in this way to popularize some cornerstone machine learning techniques. while the size of data is posing challanging questions on how to efficiently pro and they have proved successful to approach data analysis in many different areas, not only including physics, impacting

Pervasiveness of distributed cloud-based clusters, crucially providing both the computational and the storage resources necessary for processing massive datasets;

The data-driven paradigm is emerging, the internet of things (IoT) is promising an even more increasing volume of data to be digested in the upcoming years, pressing for successful strategies capable of processing at high rate (or even in real-time) massive distributed datasets.

Mastering these topics requires a mix of knowledge, skills and experience to succeed, together with high proficiency with related algorithms and their implementation. Familiarity with algorithms and ability to efficiently implement them is an essential part, and it can become even less trivial when the data requires distributed or high-performance computing.

Two schools: frequentist and Bayesian. Unfortunately, there is no general consensus among statisticians (and even among physicists)

The aim of this notes is slightly ambitious. First, thoroughly discuss at graduate level the theoretical underpinnings of traditional and modern statistical tools and their mathematical foundations, establishing the main results and developing theoretical tools. I firmly believe that it is not possible to perform a reliable data analysis without a solid knowledge of the theory behind those tools: theory is essential to understand and control the range of validity of a given method, to understand how to inspect the results, and to highlight artifacts (which are always present), and in case to develop custum solutions for the problem at hand. In the second part, we will discuss concrete examples, together with actual implementation of algorithms to handle data analysis in real examples. This will help to illustrate pitfalls, best practises, implementation strategies, libraries and platforms, etc

Trieste, March 14, 2021 *Alessandro Candolini*

Part I

PROBABILITY

This part deals with the theory of probability. The rigorous definition of probability is given, according to Kolmogorov. Then, its properties are derived, a number of recurrent common probability distributions are introduced together with tools to handle them (among them are: probability distributions, characteristic functions, moments, etc) and finally few important limit theorems are established. The theory of probability is important by its own and it is at the same time the natural prerequisite for statistics.

1 RANDOM VARIABLES

Discrete and continuous, univariate and multivariate real random variables are discussed, together with a bunch of practical tools helping to handle with them, specifically: probability density function, cumulative distribution function, moments (i. e., mean, variance, skewness, kurtosis, etc). All of this provides a more concrete, visualizable approach to probability built on top of the construction of the previous chapter by means of probability measures; it should be stressed however that the probability measures provides a more general framework. This chapter also covers functions of random variables, change of variables (i. e., what physicists are accustomed to call "propagation of errors"), etc. Generating functions — another useful tool to deal with probability distributions — are postponed to chapter 3: an entire chapter is devoted to them since their theory and their application deserve special attention. More general random variables (e. g., matrix-valued random variables) although useful are not presented here (see however chapter 10, which is devoted to a brief account of random matrix theory).

1.1 DEFINITIONS AND BASIC PROPERTIES

Let (a) $(\Omega, \Sigma, P(\cdot))$ be any probability space; (b) (E, ϵ) be any measure space; (c) $X\colon \Sigma \to E$ be any (measurable) function from Σ to E. In this context: Ω is called the "sample space", E is called "state space", X is called a "random variable", "aleatory variable" or "stochastic variable".

Whenever the range $X(\Sigma) \subseteq E$ of X has finite or countably infinite cardinality, the random variable X is called a "discrete" random variable. When the range of X is uncountably infinite instead, X is called a "continous" random variable.

In this chapter, we will be mainly interested in the case of *real*-valued (discrete or continuous) univariate (i. e., $E \subset \mathbb{R}$) or multivariate (i. e., $E \subseteq \mathbb{R}^N$, $N \in \mathbb{N}$, $N \geqslant 2$) random variables (the Borel algebra being the underlying σ-algebra in the state space). For such case, it makes sense and it proves useful to introduce the important concepts of moments, distribution functions, etc. (Those definitions are slightly different whether the random variable is discrete or continuous and univariate or multivariate, as we will see shortly.) However, the definition of random variables applies to more general settings; for example, in chapter 10 the case where E is a suitable space of matrices will be taken into account. Other scenarios are also possible (e. g., random

graphs, etc) some being relevant for statistical mechanics, machine learning, etc (see later chapters).

It is worth notice that random variables take value over the set E, nevertheless E is not the probability space itself; the random variable instead is defined as a mapping defined over a suitable σ-algebra of an underlying probability space Ω and targetting the (measure) state space E.

What is the benefit of having defined a random variable as a mapping from a probability space to the state space E instead of having E itself equipped with a suitable probability measure?

1.2 PROBABILITY DISTRIBUTION

In most applications, the underlying probability space defining a random variable remains hidden. Instead, a "distribution function" is introduced which

1.2.1 *Univariate discrete random variable*

1.2.2 *Univariate continuous random variable*

1.2.3 *Multivariate discrete random variable*

1.2.4 *Multivariate continuous random variable*

1.3 MOMENTS OF A DISTRIBUTION

1.4 TRANSFORMATIONS OF RANDOM VARIABLES

1.5 BUFFON'S NEEDLE

1.6 WHAT'S THE USE OF ALL THIS?

REFERENCES FOR CHAPTER 1

BILLINGSLEY, P., *Probability and Measure*, 3rd ed., Wiley Series in Probability and Statistics, Wiley, Hoboken, NJ 1995.

COX, R. T., "Probability, frequency and reasonable expectation", *Am. J. Phys.* 17 (1946), pp. 1-13.

FASANO, A. and S. MARMI, *Analytical Mechanics: An Introduction*, Oxford graduate texts in mathematics, Oxford University Press, Oxford 2006.

FELLER, W., *An introduction to probability theory and its applications*, 3rd ed., Wiley mathematical statistics series, Wiley, New York 1966, vol. 1.

JAYNES, E. and G. BRETTHORST, *Probability Theory. The Logic of Science*, Cambridge University Press, Cambridge 2003.

KOLMOGOROV, A. N., *Foundations of the theory of probability*, 2nd ed., Chelsea Pub. Co., New York 1956.

ROSENTHAL, J. S., *A first look at rigorous probability theory*, World Scientific, Singapore 2006.

STROMBERG, K., "The Banach-Tarski Paradox", *The American Mathematical Monthly*, 86, 3 (1979), pp. 151-161.

VAN HORN, K. S., "Constructing a logic of plausible inference: A guide to Cox's theorem", *International Journal of Approximate Reasoning*, 34 (2003), pp. 3-24.

VAN KAMPEN, N. G., *Stochastic processes in physics and chemistry*, 3rd ed., Elsevier, Amsterdam 2007.

2

CATALOG OF PROBABILITY DISTRIBUTIONS

The tools developed in chapter 1 are applied to a list of important discrete and continuous, univariate and multivariate probability distributions. These distributions are the building blocks of several probabilistic models and are relevant for later applications in statistics, so it is worth browsing through all the details and calculations.

3 LIMIT THEOREMS

3.1 INEQUALITIES

Part II

ADDITIONAL TOPICS IN PROBABILITY

The material in this part might appear to be slightly departing from the core topics of these notes, namely statistics, however much of the material is relevant for stochastic modelling.

4

STOCHASTIC PROCESSES

REFERENCES FOR CHAPTER 4

WILF, H. S., *Generatingfunctionology*, 2nd ed., Academic Press, London 1994.

5

MARKOV PROCESSES

An important class of stochastic processes is provided by the so-called Markov processes. Markov processes are ubiquitous. They are a relevant ingredient for stochastic modelling and they provide the theoretical basis of important statistical and coomputational techniques, ranging from Markov-chain Montecarlo to hidden Markov models, discussed in later parts. This chapter serves as an introduction to their definitions and first properties, and illustrates few beginning examples. The material is indispensable also for later chapters.

The following stochastic processes *need not* to be Markovian. They can be Markovian in some special cases but they do not need to be Markovian in general and allow for special treatment. For example, the simple random walk is Markovian; the self-avoiding random walk is a prototypical example of a *non*-Markovian process. For this reason, these processes will be discuss in later chapters:

- random walks;
- branching processes.

As prototypes of Markov processes we will discuss:

- Erenfest urn;
- Google™ PageRank™ ranking algorithm;
- Applications of Markov processes to shopping;
- Solution of the coupon problem via Markov chains.

In later chapters, we will discuss modern topics like

- Markov chain Montecarlo and Metropolis-Hasting algorithm in computational statistical mechanics and computational Bayesian inference;
- Hidden Markov models.

Both heavily relies on Markov processes. The stochastic calculus and the chapter on stochastic differential equation makes use of Wigner integrals which are based on Markov processes.

5.1 MARKOV PROPERTY

Intuitively speaking, Markov processes are *memoryless*: the probability of evolving from the current state to the next one does not depend on

FIGURE 5.1. Russian mathematician Andrei Andreyevich Markov (1856–1922), the father of Markov processes.

the previous history of the process. The exact mathematical implementation of this concept can look a bit technical for Markov processes defined on continuous state spaces and continuous index space, but for finite settings it should be easier to grasp. For this reason, we first define the general notion for arbitrary (continuous) processes but later we will mainly focus on the finite setting, where Markov processes are more often called "Markov chains". Markov chains will be enough for most of our purposes. General Markov processes occur however in many situations (Brownian motion, Markovian evolution of open quantum systems, etc); for this reason, we give the general definitions.

5.2 ERGODICITY

5.3 STATIONARY DISTRIBUTION

5.4 REMARKS ON NON-MARKOVIAN PROCESSES

*

5.5 THE COUPON PROBLEM REVISITED

5.6 ERENFEST URN

5.7 SHOPPING MODEL

5.8 GOOGLE™ PAGERANK™ ALGORITHM

In the early days of web search engine,

* **Van-Kampen:1998**.

Web search engines usually work in this way: (a) Web Crawling (b) Indexing (c) Ranking First of all, an automated Web crawler retrieves and stores information from the HTML markup of several many web pages. Data is analyzed and stored in a index database for use in later queries. Finally, when a query is performed by the user, the list of matching results needs to be sorted by some criteria (ranking).

The PageRank™ algorithm was has a lot of interesting mathematics behind it. In this note we will use the PageRank™ algorithm a a prototypical model and as a pathfinder to explore and to illustrate some interesting mathematics behind it. We will give also a toy implementation of the algorithm to start playing with it so to discover some of its features.

6

POISSON PROCESSES

Poisson processes and their characterization is presented.

7

RANDOM WALKS

REFERENCES FOR CHAPTER 7

VAN KAMPEN, N. G., "Remarks on Non-Markov Processes", *Brazilian Journal of Physics*, 28, 2 (June 1998).

8

BRANCHING PROCESSES

For a superb account of branching processes, see **Harris:2002**[*].

[*] .

9

GLIMPSE AT RANDOM MATRIX THEORY

10

ITÔ STOCHASTIC DIFFERENTIAL EQUATIONS

REFERENCES FOR CHAPTER 10

HARRIS, T. E., *The Theory of Branching Processes*, Dover phoenix editions, Dover Publications, New York 2002.

Part III

STATISTICAL INFERENCE

After introducing the basis of descriptive statistics, the core of frequentist and Bayesian inference are throughly studied.

Part IV

GLIMPSE AT STATISTICAL LEARNING

Machine learning techniques have been providing state-of-art performance in the analysis of massive datasets (even though some analysis at a rather heuristic level). This part deals with a rather introdoctory overview of some key algorithms rooted in machine learning and their statistical meaning.

11

A PANORAMA OF MACHINE LEARNING TECHNIQUES

The main source of material for this part is provided by:

- **Hastie.Tibshirani.ea:2009**[*]
- **Bishop:2006**[†]

11.1 WHAT'S THE USE OF ALL THIS?

[*] .
[†] .

12

ARTIFICIAL NEURAL NETWORKS

13

SUPPORT VECTOR MACHINES

13.1 WHAT'S THE USE OF ALL THIS?

REFERENCES FOR CHAPTER 13

BISHOP, C., *Pattern Recognition and Machine Learning*, Information Science and Statistics, Springer, New York 2006.

HASTIE, T., R. TIBSHIRANI, and J. FRIEDMAN, *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*, 2nd ed., Springer Series in Statistics, Springer, New York 2009.

14

GLIMPSE AT DEEP LEARNING

Part V

ADVANCED MATERIAL AND APPLICATIONS

A number of modern additional topics in statistics is covered, in particular modern computational tools (expectation-maximization algorithm, hidden-markov models).

15 KALMAN FILTERING

15.1 WHAT'S THE USE OF ALL THIS?

16 MAXIMUM ENTROPY PRINCIPLE

16.1 WHAT'S THE USE OF ALL THIS?

17 EXPECTATION-MAXIMIZATION ALGORITHM

17.1 WHAT'S THE USE OF ALL THIS?

18

HIDDEN MARKOV MODELS

18.1 WHAT'S THE USE OF ALL THIS?

19

POTTS MODELS AND SPIN GLASSES FOR IMAGE RESTORATION

19.1 what's the use of all this?

20 UNFOLDING

20.1 WHAT'S THE USE OF ALL THIS?

Part VI

APPENDIX

A

IMPLEMENTATION NOTES

It is worth summarizing few notes about implementation.

A.1 GLIMPSE AT FUNCTIONAL PROGRAMMING LANGUAGES

Functional programming languages has gained attraction over imperative ones in recent years due to their ability to handle effortlessly and in robust way concurrency and multithreading/multicore computing, making them particularly suitable to distributed frameworks.

Historically, concurrent programming was not the original motivation for functional programming. After the early functional-flavored language Lisp and the development of APL in the early 1960s by Iverson, functional programming was eventually formalized by John Bakus (the father of Fortran) in his cornerstone 1977 Turing Award Lecture *Can Programming Be Liberated From the von Neumann Style? A Functional Style and its Algebra of Programs*, where he presented his FP programming language and emphasizes how liberating a programming language from the traditional imperative style would have lead to a programming language closer to mathematical abstractions.

For several years, functional programming was relegated mainly as an academic tool. Nevertheless, several powerful functional programming languages have been developed in the last decades. It is worth mentioning at this point: Haskell, Erlang, OCam, Closure. Scala was build

A.2 REACTIVE PARADIGM

A.3 PYTHON

A.4 IPYTHON AND OTHER NOTEBOOKS

A.5 SCALA

A.6 CERN ROOT FOR HEP

In high-energy physics

A.7 DISTRIBUTED PROGRAMMING: APACHE SPARK

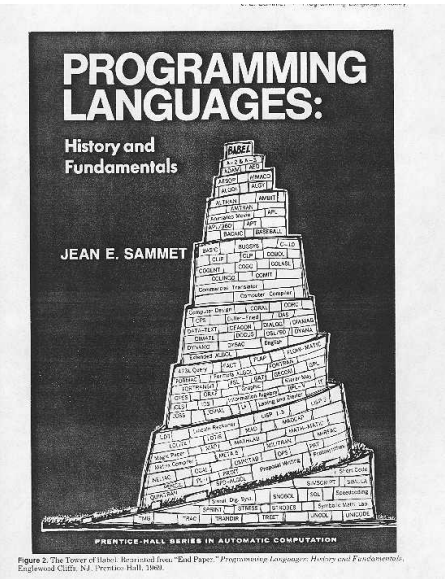


FIGURE A.1. The Babel's Tower of programming languages, taken from the cover of J. Sammet's *Programming Languages: History and Fundamentals* (1969). Being a 1969 book, the picture misses several important additions, nevertheless it is still representative of the difficulty in choosing a suitable programming language.

BIBLIOGRAPHY

REFERENCES FOR CHAPTER 1

BILLINGSLEY, P., *Probability and Measure*, 3rd ed., Wiley Series in Probability and Statistics, Wiley, Hoboken, NJ 1995.

COX, R. T., "Probability, frequency and reasonable expectation", *Am. J. Phys.* 17 (1946), pp. 1-13.

FASANO, A. and S. MARMI, *Analytical Mechanics: An Introduction*, Oxford graduate texts in mathematics, Oxford University Press, Oxford 2006.

FELLER, W., *An introduction to probability theory and its applications*, 3rd ed., Wiley mathematical statistics series, Wiley, New York 1966, vol. 1.

JAYNES, E. and G. BRETTHORST, *Probability Theory. The Logic of Science*, Cambridge University Press, Cambridge 2003.

KOLMOGOROV, A. N., *Foundations of the theory of probability*, 2nd ed., Chelsea Pub. Co., New York 1956.

ROSENTHAL, J. S., *A first look at rigorous probability theory*, World Scientific, Singapore 2006.

STROMBERG, K., "The Banach-Tarski Paradox", *The American Mathematical Monthly*, 86, 3 (1979), pp. 151-161.

VAN HORN, K. S., "Constructing a logic of plausible inference: A guide to Cox's theorem", *International Journal of Approximate Reasoning*, 34 (2003), pp. 3-24.

VAN KAMPEN, N. G., *Stochastic processes in physics and chemistry*, 3rd ed., Elsevier, Amsterdam 2007.

REFERENCES FOR CHAPTER 4

WILF, H. S., *Generatingfunctionology*, 2nd ed., Academic Press, London 1994.

REFERENCES FOR CHAPTER 7

VAN KAMPEN, N. G., "Remarks on Non-Markov Processes", *Brazilian Journal of Physics*, 28, 2 (June 1998).

REFERENCES FOR CHAPTER 10

HARRIS, T. E., *The Theory of Branching Processes*, Dover phoenix editions, Dover Publications, New York 2002.

REFERENCES FOR CHAPTER 13

BISHOP, C., *Pattern Recognition and Machine Learning*, Information Science and Statistics, Springer, New York 2006.

HASTIE, T., R. TIBSHIRANI, and J. FRIEDMAN, *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*, 2nd ed., Springer Series in Statistics, Springer, New York 2009.