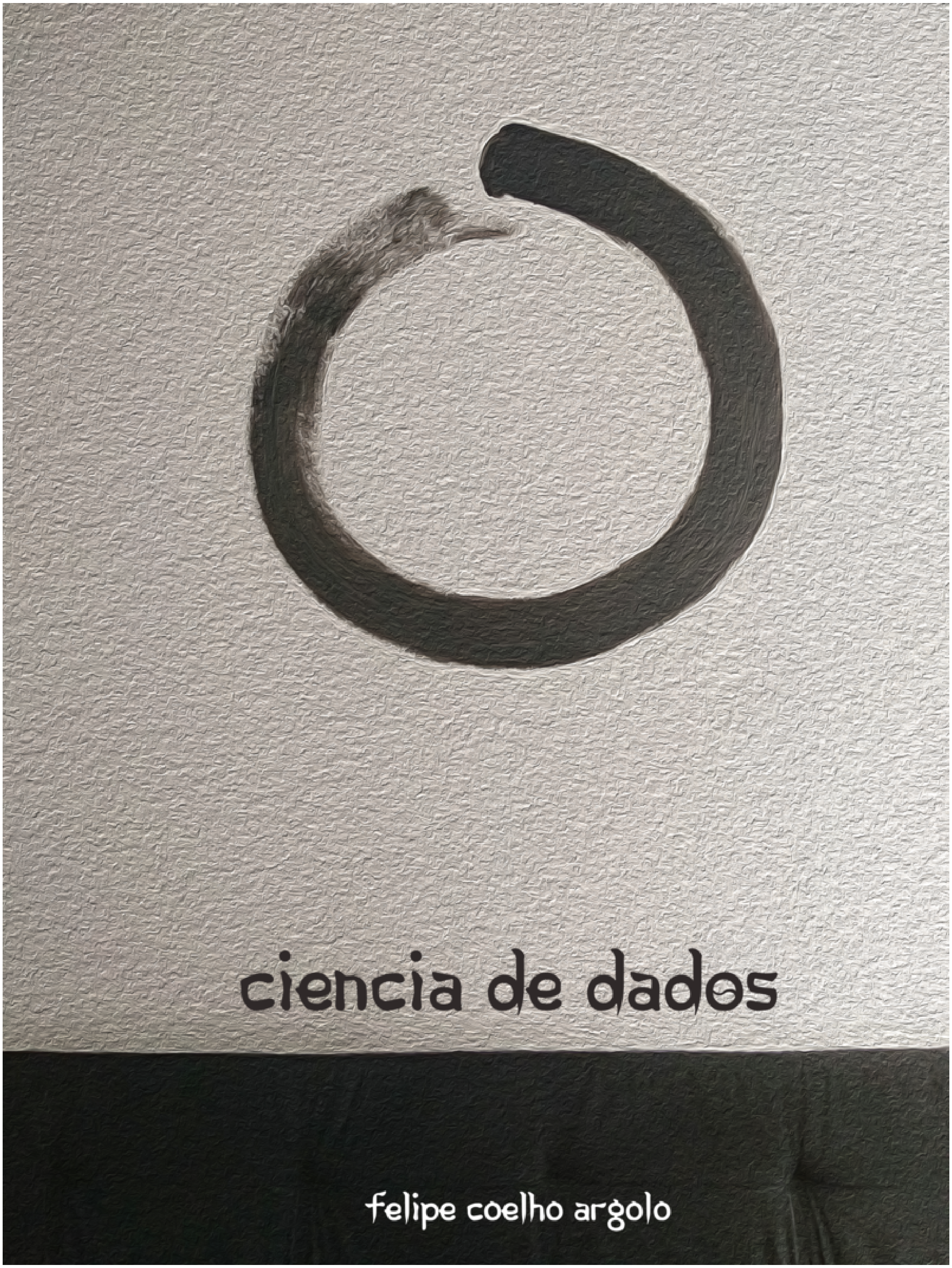

ciencia de dados

felipe coelho argolo

Data Science

Philosophers guide with software applications

Felipe Coelho Argolo felipe.c.argolo@protonmail.com

São Paulo, 21th of June, 2019

Official page: https://http://www.leanpub.com/fargolo

Volume 1

Preface

Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful
George Box & Norman R. Draper, Empirical Model-Building and Response Surfaces

In recent years, the terms *artificial intelligence*, *machine learning*, *big data*, and *data science* receive much attention due to unprecedented applied results. Advances in natural language processing, computer vision, and predictive algorithms were quickly applied by engineers and researchers in finance, industry and research. A comprehensive description of the techniques developed can easily reach over 1,000 pages of succinct text, such as Goodfellow's classic 'Deep Learning' (Adaptive Computation and Machine Learning), Bengio and Courville. Another work of similar scope and size is Simon Haykin's "Neural networks and learning machines". Numerous online courses and videotapes are produced and made available by prestigious institutions (e.g. Oxford Comprehensive Course in Deep Learning: https://www.youtube.com/watch?v=PlhFWT7vAEw).

The content topics are usually approached in a '* bottom-up _' manner. A notion of the field is constructed through focused study of models: specific courses for_ time series, clustering, neural networks* or tools (eg R, Julia, Python, Stan, Matlab . . .). It works well as a natural course in engineering and science courses.

Top down
This text visits themes in reverse order (*top-down*). The models are contextualized as tools in the exploration of a script whose main axis is centered on **philosophy of the sciences**. The formulations arise as assets to questions about natural phenomena in biology (statistical tests), psychology (factorial analysis), public health / economy (correlation, regression and causality) and neurosciences (perceptron and neural networks).

The *first chapter* accompanies Charles Darwin in the Galapagos. Darwin waited 20 years between the conception of the theory and its publication. He worked tirelessly to investigate whether his impressions were not false. This chapter illustrates how the hypothetical-deductive rational works to study scientific hypotheses. Student's t test is applied for comparing the beaks of birds in Galapagos. It addresses the relationship between empirical sciences, the Central Limit Theorem, the Normal Distribution, and the t distribution.

The second chapter highlights the descriptive and predictive role of theories. In addition to testing hypotheses, we create models for the relationships between measures. From Archimedes' studies on levers, we will learn linear correlations (Pearson‘s ρ) and effect size (*Cohen s D*). Non-parametric alternatives to previous Spearman procedures (ρ and Mann-Whitney U test) are also introduced.
We use regression to make predictions and estimate parameters using *closed forms*, analyzing the equations of the model analytically to achieve the best result.

The *third chapter* introduces the mathematical implementation of a comprehensive philosophical paradigm for Judea Pearl‘s **causal models**. Considering many variables (multivariate analysis), graphs are the basic abstraction for relating concepts. We will study multiple linear regression, collinearity, mediation and moderation. Also, factorial analysis, principal component analysis (PCA) and its generalization in structural equations (SEM).

The *fourth chapter* introduces neural networks. We begin with the biological inspiration of the artificial neurons and the first intelligent machine in history: the *Mark I Perceptron.* We encode a virtual Mark I, which uses a new way of estimating parameters: *gradient descent.* Instead of using a closed formula solution, we use derivatives to 'walk' toward the minimum progressively.

Neural networks expand the power of a neuron, using multiple nodes to construct complex predictive systems. Deep networks include successive layers, allowing transformations in sequence to solve more general classes of problems. We will understand how neurons can propagate errors to others by optimizing gradients in conjunction with the backpropagation mechanism. We will also encode a neural network, Mark II.

The *fifth chapter* contrasts the two main schools in interpretation of probability: the **frequentist** and the **bayesian** ones. The context is given by alternatives to the hypothetical-deductive method: *Carnap* demonstrates the logical difficulties implied in falsifiability, while Feyerabend proposes an epistemological anarchy supported by historical facts and W. van Quine paints a system intertwining theories, hypotheses

and observations. We reapproach previous examples using *Stan* for Bayesian inference.
We then explore a third way of estimating parameters. Without closed formulas nor derivatives (*gradient descent*), we use the power of stochastic simulations (*Markov Chain Monte Carlo*).

Summary

  - Intuitions: prior, likelihood, posterior and marginal probabilities
  - Comparison of samples with normal distribution
  - Linear Correlation
- Estimators and Methods Markov Chain Monte Carlo
  - Closed solutions, Gradient Descent and MCMC

Requirements for this text

Rudiments in probability, statistics and calculation are sufficient to understand almost all the examples. The programs use syntax similar to the math presented in the text, hence little familiarity with programming is not a barrier. Chapter 0 addresses this.

All examples can be reproduced using free software.

Recommended Reading:

Philosophy and Science

- Surely You're Joking, Mr. Feynman
- The World Haunted by Demons - Carl Sagan
- The logic of scientific research - K. Popper
- The Structure of Scientific Revolutions - Thomas Kuhn
- Against the Method - Paul Feyerabend
- Two dogmas of empiricism - Willard van Quine
- Stanford Encyclopedia of Philosophy (https://plato.stanford.edu/) entries are valuable assets. Themes not mentioned above:
  - Wittgenstein notion of language games

  - Imre Lakatos conceptualization of Popper‘s theory

Neurosciences

- Principles of neural science - Eric Kandel

Mathematics / computation

- What is mathematics - Courant & Robbins
- Better Explained (https://betterexplained.com/)
- http://material.curso-r.com/
- R Graphics Cookbook
- R Inferno
- Learn you a Haskell for Great Good
- Layered Grammar of Graphics - Hadley Wickham.
- Algorithms unlocked
- Online: Statsexchange, stackoverflow, mathexchange, cross-validated.

Machine Learning

- An Introduction to Statistical Learning: with Applications in R
- Neural Networks and Learning Machines - Simon Haykin
- Stanford (computer vision): http://cs231n.stanford.edu/
- Oxford 2015 (Deep learning): (https://www.youtube.com/watch?v=dV80NAlEins&list=PLE6Wd9FR--EfW8dtjAuPoTuPcqmOV53Fu)

Acknowledgements

To my family, Suzana, Paulo, Isaac e Chris. Friends Gabriel, Joao, Guilherme, Wei.

Professors Carla Daltro, Anibal Neto, Lucas Quarantini, Luis Correia, Rodrigo Bressan, Ary Gadelha.

Colleagues Fatori, Luccas, Macedo, Walter, Sato, André, Hiroshi, Lais, Luci, Davi, Oddone, Jones, n3k00n3 (Fernando), Loli (Lorena).

For comments, critique, suggestions or just saying *hello*: felipe.c.argolo@protonmail.com.