Compact Textbooks in Mathematics

Compact Textbooks in Mathematics

This textbook series presents concise introductions to current topics in mathematics and mainly addresses advanced undergraduates and master students. The concept is to offer small books covering subject matter equivalent to 2- or 3-hour lectures or seminars which are also suitable for self-study. The books provide students and teachers with new perspectives and novel approaches. They feature examples and exercises to illustrate key concepts and applications of the theoretical contents. The series also includes textbooks specifically speaking to the needs of students from other disciplines such as physics, computer science, engineering, life sciences, finance.

More information about this series at http://www.springer.com/series/11225

Victor M. Panaretos

Statistics for Mathematicians

A Rigorous First Course

Birkhäuser

Victor M. Panaretos
Institute of Mathematics
EPFL
Lausanne, Switzerland

ISSN 2296-4568 ISSN 2296-455X (electronic)
Compact Textbooks in Mathematics
ISBN 978-3-319-28339-5 ISBN 978-3-319-28341-8 (eBook)
DOI 10.1007/978-3-319-28341-8

Library of Congress Control Number: 2016940379

Mathematics Subject Classification (2010): 62-XX

© Springer International Publishing Switzerland 2016
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

This book is published under the trade name Birkhäuser
The registered company is Springer International Publishing AG Switzerland
(www.birkhauser-science.com)

In memory of David A. Freedman,
master of clarity

Preface

This book is intended as a text for Mathematics students taking their first course in Statistics, and grew out of my second-year course for mathematics undergraduates at EPFL. It is a book on "Statistics for Mathematicians" rather than on "Mathematical Statistics": the intent is not to focus on the deeper mathematical/theoretical aspects of the subject but rather to provide an introduction to the basic notions tailored to the mindset and tastes of the Mathematics student. Mathematics students are sometimes put off by the informal nature of first courses in Statistics, since many results are usually stated without proof or are accompanied by heuristic sketches of proofs. Another risk may be that of "intellectual entropy", when too many (and diverse) topics are covered in a single course, risking the impression of Statistics as a collection of recipes lacking natural connection. This book can be used as a basis for an elementary semester-long first course in Statistics that presents the basic ideas of one-parameter inference in a coherent manner, while making essentially no sacrifices on matters of rigour. It is meant to be compact, so as to be realistic to be covered in full during a single semester, and yet hopefully attract mathematics students to pursuing further elective courses in Statistics. In more detail, the three main tasks this text sets out to address are as follows.

(1) To provide a rigorous yet elementary course The effort is to prove essentially all the results rigorously. These results include some of the most central results such as the asymptotics of maximum likelihood, optimality in testing, asymptotics of likelihood ratio tests, and optimality results regarding confidence intervals. It also contains detailed proofs of some elementary results that are rarely worked out in detail in elementary texts (for instance, the derivation of the distribution of the t statistic). The only results not proven in the main text are some background results in probability and analysis. In the case of the probabilistic results, detailed proofs *are* in fact given in the appendix, and the proofs are still at an elementary level. These include results such as the continuous mapping theorem, Slutsky's theorem, the (third moment) central limit theorem, and results pertaining to moment generating functions. The analytic results not proven are Taylor's formula and the univariate inverse function theorem. These are stated in the appendix, where precise references are also provided for their proofs. In principle, thus, the course only requires students to have taken a first course in ϵ/δ-level analysis (including sequences, convergence, series, multivariable differentiation and Riemann integration, and Taylor's

formula) and a first course in probability (including basic operations on events and the corresponding probability calculus, discrete and continuous random variables, joint/conditional/marginal distributions, and expectation/variance/covariance). A succinct fact sheet on all the probabilistic prerequisites is provided in the appendix, for easy reference.

(2) To provide a conceptually compact course, with a firm sense of direction The entire book can realistically be covered in full during the course of a semester, and it is also realistic for the students to solve all the exercises during the same period of study (a solution manual is available upon request for instructors). I have reduced the number of topics covered in order to be able to have the minimal number of topics that can be covered during a semester course without compromising on the mathematics, while still providing an overview of the main ideas of statistical inference. The course covers the basics of exponential families, exploratory data analysis, sampling, estimation, testing, and confidence intervals. It's true that the book does not tell the whole story and avoids detailed discussions of all the possible complications and variants in each section. However, I believe that the topics covered give a firm basis for the students to build on, and every attempt has been made for the story it tells to flow naturally, without giving the appearance of a collection of techniques. There is extensive cross-referencing of the material, illustrating how the different results are tied together, and an effort to develop the material in a "linear fashion", explaining why one is doing whatever they are doing at every point, and what the ultimate purpose is. No result is mentioned in vain (any results presented are subsequently used), and results are accompanied by substantial motivation and discussion. References made to results are always accompanied by the number of the said result, along with the page number in the book which allows for easy reference and self-study.

(3) To provide a course that is not on "Mathematical Statistics" but rather on "Statistics for Mathematicians" The audience is primarily intended to be undergraduate mathematicians, whom I hope to attract into Statistics rather than statisticians to whom I might want to introduce the more mathematical aspects of Statistics. Therefore, the course is not primarily intended to be a course in statistical theory. Rather, it is intended to be an entry-level course in statistical inference, presented in a way that would be more receptive by an audience comprised of mathematicians. Therefore, the discussion of different topics and the style and considerations are adapted to such an audience. For example, optimality, whenever discussed, is not presented as an end in itself but rather as a means of motivating methodology (the idea being that mathematicians would be motivated by "best" results more than by heuristics).

The means to balance the requirement of an elementary yet rigorous text was to adopt the use of the exponential family of distributions throughout (rather than aiming for full generality). This is of course a restriction, but in some ways not a major one, since most of the examples treated in elementary textbooks *are*, in fact, exponential families. Focusing on exponential families not only allows for elementary proofs using basic analysis and probability but also allows for the

statements of the theorems and the required conditions to be simple and intuitive. Whenever results do hold more generally, this is remarked as a side note. A more detailed description of the structure of the text, and the progression of topics, can be found in the "Brief Overview" Section (p. 1).

The main concessions that regrettably had to be made in terms of coverage pertain to regression and the Bayesian paradigm, and this deserves an apology. The textbook is based on the first Statistics course that mathematics students take at EPFL, but this course is also the only *compulsory* course in Statistics. It may thus well be their last (though hopefully the book will convince them otherwise). In this case there is a dilemma. Does one strive to include as many topics as possible, so that the student be well equipped in the future in case this is all the Statistics they will ever see? Or does one try to cover a minimally sufficient number of topics as clearly and completely as possible, hoping that at least these topics will stick to mind? I opted for the second approach, as my impression is that adding more topics does not guarantee that these topics will in fact be remembered (in fact, a student having only taken a single Statistics course and finding themselves needing Statistics later will almost certainly have to do further reading anyway) and because this approach is more in line with the effort to produce a course with low conceptual entropy. For instance, notions such as p-values and confidence intervals are quite subtle to understand upon a first encounter (avoiding flawed interpretations such as "the probability that H_0 is valid" or "the probability that the parameter falls in the interval is 95%"). When the student does not already have a solid grasp, it may be unsettling—or worse still confusing—to suddenly switch things around.

In writing the book, and preparing examples and exercises, I have drawn inspiration from many excellent textbooks that have stood the test of time (but also more recent online resources, including Wikipedia and mathstackexchange). In doing this, I tried to balance the rigour found in advanced textbooks focusing on Mathematical Statistics, with the more accessible style of entry-level textbooks focusing on the basics of statistical inference. The former category includes Lehmann and Casella [15], Lehmann and Romano [16], Cox and Hinkley [6], Bickel and Doksum [1], Schervish [22], Shao [23], and Young and Smith [26], and the latter category includes Rice [19], Hogg and Tanis [13], Hogg and Craig [12], and Silvey [24] (the last one perhaps bordering with the first category). The book by Knight [14] strikes a very nice balance between the two objectives, though still at a level higher than the present text aims, and has also been an important source of inspiration and exercises/examples. More texts striking a good balance and including a more comprehensive list of topics than the present one (but still not including several proofs) include Casella and Berger [4], Davison [9], and Wasserman [25]. The necessary probability background for the present text is covered quite nicely in the first three chapters of Knight [14], but of course there are several texts devoted specifically to elementary probability (i.e. non-measure theoretic probability) that would suffice (e.g. Blitzstein and Hwang [3], Dalang and Conus [8] (in French), Grimmett and Welsh [11], Pitman [18], and Ross [20]). As mentioned earlier, Sect. A.1 contains a quick overview of the main prerequisites, for ease of reference.

While the main audience for the book will be instructors and students in mathematics undergraduate programmes, the textbook could still be used for programmes of study with substantial mathematical content, for instance, students of physics, economics, computer science, and engineering programmes looking for a more formal coverage of one-parameter inference. After all, to think like a mathematician is to think rigorously, regardless of the subject matter at hand.

In closing, I would like to express my gratitude to my PhD students and my undergraduate students whose meticulous comments and suggestions helped improve earlier drafts. Marie-Hélène Descary, Mikael Kuusela, Valentina Masarotto, Matthieu Simeoni, and Yoav Zemel provided extensive feedback, suggestions on exercises, and help with proofreading and layout. I especially enjoyed chatting with Yoav Zemel about how to best tiptoe around measure theory in the proofs of some more delicate results in the appendix (while remaining fully rigorous). I am also very thankful to two anonymous reviewers, who read a first version of the book and gave constructive and encouraging feedback. Any remaining glitches are, of course, my own. Finally, I would like to thank Veronika Rosteck and Springer/Birkhäuser for our pleasant collaboration.

Lausanne, Switzerland Victor M. Panaretos
October 2015

Contents

Brief Overview

In a general sense, one can describe Statistics as the mathematical discipline whose purpose is to

use empirical data generated by a random phenomenon, in order to

make inferences about some deterministic characteristics of the phenomenon

while simultaneously quantifying the uncertainty inherent in these inferences.

Let's take a step back and consider the different elements of this description. What is a random phenomenon? We can think of a random phenomenon as a system or process whose outcome X is uncertain. This means that, even if we know every aspect of this system or process, we cannot perfectly predict its outcome X. Mathematically, such phenomena are formalised via the theory of probability: the outcome X is a random variable, and the model that describes the phenomenon is the probability distribution function $F(x) = \mathbb{P}[X \leq x]$ of this random variable. Now there may be a characteristic θ of this phenomenon that influences the probabilities associated with the outcome of X. Such a characteristic is called a parameter. Since the probability of $\{X \leq x\}$ is influenced by θ, the function $F(x)$ must be a function of θ, so we write it as $F(x;\theta) = \mathbb{P}_\theta[X \leq x]$.

If we know the functional form of $F(x;\theta)$, and the true value of θ, we can then calculate the probability $\mathbb{P}_\theta[X \leq x] = F(x;\theta)$ for any possible outcome x. Statistics deals with the inverse problem: suppose that we know the precise functional form of $F(x;\theta)$, but do not know which is the true θ. If we have an outcome x (a realisation of X), is it possible to say something useful about θ? It seems that we should be able to do so. Since θ influences what outcomes are most probable, then knowing an outcome should give us information on which θ are plausible. The topic of this text will be how exactly to make this connection rigorous and show how to exploit it in order to (a) make the best possible use of our data x to better inform ourselves about θ and (b) understand how certain we can be about our inferences on θ for the given data x. In summary, our framework is as follows:

1. There is a distribution $F(x;\theta)$ depending on an unknown $\theta \in \mathbb{R}^p$.
2. We observe the realisation of n independent identically distributed random variables $X_1, \ldots, X_n$ that follow this distribution.

3. We wish to use our n observations (the realisations of $X_1, \ldots, X_n$) in order to make statements about the true value of θ and to quantify the uncertainty associated with those statements.

At first glance, this framework may seem restrictive. Indeed, it represents a significant simplification over the much broader framework where one can develop statistical methodology. For example, in general, the unknown parameter of interest θ might not be an element of $\mathbb{R}^p$, but an element of a more general mathematical space (a space of functions, for instance). Also the data $(X_1, \ldots, X_n)$ could be dependent; they could themselves be vectors, or functions, or some more general mathematical object.

However, some of the key ideas that statisticians employ in order to attack these more general situations are already present in the simpler scenario that we will consider in this text. In fact, many highly more complex situations can often be reduced to this simpler case by a careful use of mathematics (for example, a real function can be identified with a vector in $\mathbb{R}^p$ when represented by its basis coefficients in some basis expansion, a dependent collection of random variables might in fact be approximately independent, and so on). In a sense, the framework we will consider here is the simplest non-trivial case that nevertheless contains the germs of generality.

Following is an overview of the contents of this text:

1. In Chap. 1, we will review the different types of probability models that we will construct statistical methods for. We will try to understand what situations they are suitable for, and what are some of their key properties. We will also try to find a unifying framework in which we can describe several of these models at once: instead of developing results separately for each model, we will try to give an abstract description of some key common characteristics that will be useful for obtaining general results. At the end of the chapter, we will consider the problem of how to choose a type of model, whether by first principles or by means of exploratory data analysis using numerical and graphical summaries.
2. In Chap. 2, we will develop the relevant concepts and probabilistic results that are needed in order to study the problem of sampling from probability models. We will probe the behaviour of the random sample, and how this relates to the original model, and what aspects of a sample are important for the purposes of statistical inference. An important focus will be to describe the probabilistic behaviour of functions of a sample. That is, given a sample $X_1, \ldots, X_n$ from a distribution F, what is the distribution of $g(X_1, \ldots, X_n)$ for some function g? The reason we will do this is simple: all that we have available to do statistics is the sample, so anything we do will be a function of the sample!
3. Once we know what probability models we wish to consider, and how to handle samples from probability models, we will turn to the most basic statistical inference question one can ask: given a sample $X_1, \ldots, X_n$ from a distribution F_θ that depends on an unknown parameter θ, construct an estimator: a function of the sample whose purpose is to estimate θ. We will consider how to formalise

the quality of such an estimator in terms of quantifying its accuracy, and what are methods for constructing good estimators (for example, are there optimal methods?).

4. Chapter 4 deals with a somewhat different problem. Instead of trying to guess which θ was the one that generated the observed sample $X_1, \ldots, X_n$, we will attempt to answer the following question: given a candidate value θ_0 for θ (or some candidate values forming a set Θ_0), decide on the basis of the sample $X_1, \ldots, X_n$ whether this value (or set of values) is good guess for the true θ. An important part of the chapter will be devoted to making formal what we mean by candidate values, good guesses (and bad guesses), and whether there are optimal strategies to do so. We will also be considering how to quantify the accuracy of our decisions.
5. Finally, in Chap. 5, we will deal with the third of the basic trio of problems of statistical inference: confidence intervals. Roughly speaking, instead of trying to estimate the precise value of θ that generated our sample $X_1, \ldots, X_n$, we wish to provide a whole range of values in the form of some interval, which will very likely contain the true parameter θ. This chapter will formalise this notion and consider how we can construct "small" regions that have high probability of covering the true parameter θ. We will, in fact, see that the problem of constructing confidence intervals is very closely connected both with the problem of point estimation and with the problem of hypothesis testing.