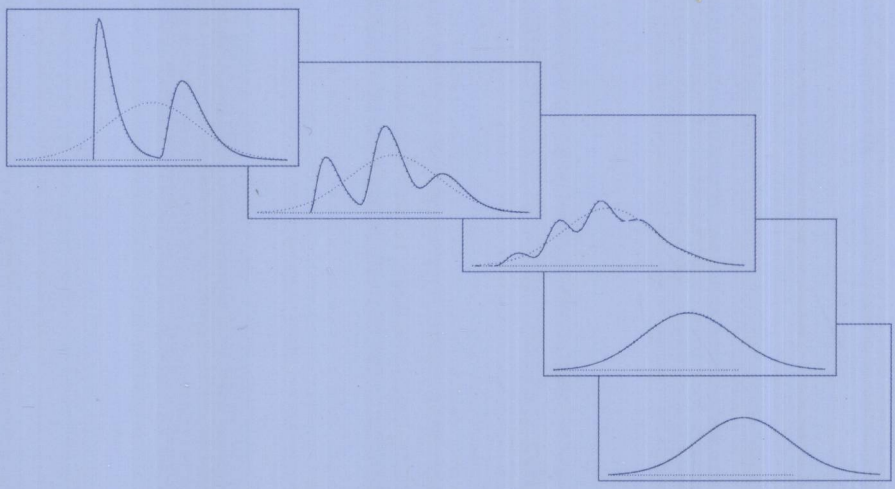
SPRINGER TEXTS IN STATISTICS

A Modern Introduction to Probability and Statistics

Understanding Why and How



F.M. Dekking
C. Kraaikamp
H.P. Lopuhaä
L.E. Meester

Springer

F.M. Dekking C. Kraaikamp
H.P. Lopuhaä L.E. Meester

A Modern Introduction to Probability and Statistics

Understanding Why and How

With 120 Figures

Springer

Frederik Michel Dekking
Cornelis Kraaikamp
Hendrik Paul Lopuhaä
Ludolf Erwin Meester
Delft Institute of Applied Mathematics
Delft University of Technology
Mekelweg 4
2628 CD Delft
The Netherlands

British Library Cataloguing in Publication Data
A modern introduction to probability and statistics. —
(Springer texts in statistics)
1. Probabilities 2. Mathematical statistics
I. Dekking, F. M.
519.2
ISBN 1852338962

Library of Congress Cataloging-in-Publication Data
A modern introduction to probability and statistics : understanding why and how / F.M. Dekking ... [et al.].
p. cm. — (Springer texts in statistics)
Includes bibliographical references and index.
ISBN 1-85233-896-2
1. Probabilities—Textbooks. 2. Mathematical statistics—Textbooks. I. Dekking, F.M. II. Series.
QA273.M645 2005
519.2—dc22 2004057700

ISBN-10: 1-85233-896-2
ISBN-13: 978-1-85233-896-1

Springer Science+Business Media
springeronline.com

Printed in the United States of America
12/3830/543210 Printed on acid-free paper SPIN 10943403

Springer Texts in Statistics

Advisors:
George Casella Stephen Fienberg Ingram Olkin

Springer Texts in Statistics

Alfred: Elements of Statistics for the Life and Social Sciences
Berger: Introduction to Probability and Stochastic Processes, Second Edition
Bilodeau and Brenner: Theory of Multivariate Statistics
Blom: Probability and Statistics: Theory and Applications
Brockwell and Davis: An Introduction to Time Series and Forecasting
Carmona: Statistical Analysis of Financial Data in S-Plus
Chow and Teicher: Probability Theory:
Independence, Interchangeability,
Martingales, Third Edition
Christensen: Advanced Linear Modeling: Multivariate, Times Series, and Spatial Data; Nonparametic Regression and Response Surface Maximization, Second Edition
Christensen: Log-Linear Models and Logistic Regression, Second Edition
Christensen: Plane Answers to Complex Questions: The Theory of Linear Models, Second Edition
Creighton: A First Course in Probability Models and Statistical Inference
Davis: Statistical Methods for the Analysis of Repeated Measurements
Dean and Voss: Design and Analysis of Experiments
Dekking et al: A Modern Introduction to Probability and Statistics: Understanding Why and How
du Toit, Steyn, and Stumpf: Graphical Exploratory Data Analysis
Durrett: Essential of Stochastic Processes
Edwards: Introduction to Graphical Modeling, Second Edition
Everitt: An R and S-PLUS® Companion to Multivariate Analysis
Finkelstein and Levin: Statistics for Lawyers
Flury: A First Course in Multivariate Statistics
Heiberger and Holland: Statistical Analysis and Data Display: An Intermediate Course with Examples in S-PLUS, R, and SAS
Jobson: Applied Multivariate Data Analysis, Volume I: Regression and Experimental Design
Jobson: Applied Multivariate Data Analysis, Volume II: Categorical and Multivariate Methods
Kalbfleisch: Probability and Statistical Inference, Volume I: Probability, Second Edition
Kalbfleisch: Probability and Statistical Inference, Volume II: Statistical Interference, Second Edition
Karr: Probability
Keyfitz: Applied Mathematical Demography, Second Edition
Kiefer: Introduction to Statistical Inference
Kokoska and Nevison: Statistical Tables and Formulae

(continued after index)

Preface

Probability and statistics are fascinating subjects on the interface between mathematics and applied sciences that help us understand and solve practical problems. We believe that you, by learning how stochastic methods come about and why they work, will be able to understand the meaning of statistical statements as well as judge the quality of their content, when facing such problems on your own. Our philosophy is one of *how* and *why*: instead of just presenting stochastic methods as cookbook recipes, we prefer to explain the principles behind them.

In this book you will find the basics of probability theory and statistics. In addition, there are several topics that go somewhat beyond the basics but that ought to be present in an introductory course: simulation, the Poisson process, the law of large numbers, and the central limit theorem. Computers have brought many changes in statistics. In particular, the bootstrap has earned its place. It provides the possibility to derive confidence intervals and perform tests of hypotheses where traditional (normal approximation or large sample) methods are inappropriate. It is a modern useful tool one should learn about, we believe.

Examples and datasets in this book are mostly from real-life situations, at least that is what we looked for in illustrations of the material. Anybody who has inspected datasets with the purpose of using them as elementary examples knows that this is hard: on the one hand, you do not want to boldly state assumptions that are clearly not satisfied; on the other hand, long explanations concerning side issues distract from the main points. We hope that we found a good middle way.

A first course in calculus is needed as a prerequisite for this book. In addition to high-school algebra, some infinite series are used (exponential, geometric). Integration and differentiation are the most important skills, mainly concerning one variable (the exceptions, two dimensional integrals, are encountered in Chapters 9–11). Although the mathematics is kept to a minimum, we strived

to be mathematically correct throughout the book. With respect to probability and statistics the book is self-contained.

The book is aimed at undergraduate engineering students, and students from more business-oriented studies (who may gloss over some of the more mathematically oriented parts). At our own university we also use it for students in applied mathematics (where we put a little more emphasis on the math and add topics like combinatorics, conditional expectations, and generating functions). It is designed for a one-semester course: on average two hours in class per chapter, the first for a lecture, the second doing exercises. The material is also well-suited for self-study, as we know from experience.

We have divided attention about evenly between probability and statistics. The very first chapter is a sampler with differently flavored introductory examples, ranging from scientific success stories to a controversial puzzle. Topics that follow are elementary probability theory, simulation, joint distributions, the law of large numbers, the central limit theorem, statistical modeling (informal: why and how we can draw inference from data), data analysis, the bootstrap, estimation, simple linear regression, confidence intervals, and hypothesis testing. Instead of a few chapters with a long list of discrete and continuous distributions, with an enumeration of the important attributes of each, we introduce a few distributions when presenting the concepts and the others where they arise (more) naturally. A list of distributions and their characteristics is found in Appendix A.

With the exception of the first one, chapters in this book consist of three main parts. First, about four sections discussing new material, interspersed with a handful of so-called Quick exercises. Working these—two-or-three-minute—exercises should help to master the material and provide a break from reading to do something more active. On about two dozen occasions you will find indented paragraphs labeled *Remark*, where we felt the need to discuss more mathematical details or background material. These remarks can be skipped without loss of continuity; in most cases they require a bit more mathematical maturity. Whenever persons are introduced in examples we have determined their sex by looking at the chapter number and applying the rule "He is odd, she is even." Solutions to the quick exercises are found in the second to last section of each chapter.

The last section of each chapter is devoted to exercises, on average thirteen per chapter. For about half of the exercises, answers are given in Appendix C, and for half of these, full solutions in Appendix D. Exercises with both a short answer and a full solution are marked with ⊞ and those with only a short answer are marked with ⊡ (when more appropriate, for example, in "Show that ..." exercises, the short answer provides a hint to the key step). Typically, the section starts with some easy exercises and the order of the material in the chapter is more or less respected. More challenging exercises are found at the end.

Springer Texts in Statistics *(continued from page II)*

Kokoska and Nevison: Statistical Tables and Formulae
Kulkarni: Modeling, Analysis, Design, and Control of Stochastic Systems
Lange: Applied Probability
Lange: Optimization
Lehmann: Elements of Large-Sample Theory
Lehmann and Romano: Testing Statistical Hypotheses, Third Edition
Lehmann and Casella: Theory of Point Estimation, Second Edition
Lindman: Analysis of Variance in Experimental Design
Lindsey: Applying Generalized Linear Models
Madansky: Prescriptions for Working Statisticians
McPherson: Applying and Interpreting Statistics: A Comprehensive Guide, Second Edition
Mueller: Basic Principles of Structural Equation Modeling: An Introduction to LISREL and EQS
Nguyen and Rogers: Fundamentals of Mathematical Statistics, Volume I: Probability for Statistics
Nguyen and Rogers: Fundamentals of Mathematical Statistics, Volume II: Statistical Inference
Noether: Introduction to Statistics: The Nonparametric Way
Nolan and Speed: Stat Labs: Mathematical Statistics Through Applications
Peters: Counting for Something: Statistical Principles and Personalities
Pfeiffer: Probability for Applications
Pitman: Probability
Rawlings, Pantula, and Dickey: Applied Regression Analysis
Robert: The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation, Second Edition
Robert and Casella: Monte Carlo Statistical Methods, Second Edition
Rose and Smith: Mathematical Statistics with *Mathematica*
Ruppert: Statistics and Finance: An Introduction
Santner and Duffy: The Statistical Analysis of Discrete Data
Saville and Wood: Statistical Methods: The Geometric Approach
Sen and Srivastava: Regressions Analysis: Theory, Methods, and Applications
Shao: Mathematical Statistics, Second Edition
Shorack: Probability for Statisticians
Shumway and Stoffe: Time Series Analysis and Its Applications
Simonoff: Analyzing Categorical Data
Terrell: Mathematical Statistics: A Unified Introduction
Timm: Applied Multivariate Analysis
Toutenburg: Statistical Analysis of Designed Experiments, Second Edition
Wasserman: All of Statistics: A Concise Course in Statistical Inference
Whittle: Probability via Expectation, Fourth Edition
Zacks: Introduction to Reliability Analysis: Probability Models and Statistical Methods

Contents

1

Why probability and statistics?

Is everything on this planet determined by randomness? This question is open to philosophical debate. What is certain is that every day thousands and thousands of engineers, scientists, business persons, manufacturers, and others are using tools from probability and statistics.

The theory and practice of probability and statistics were developed during the last century and are still actively being refined and extended. In this book we will introduce the basic notions and ideas, and in this first chapter we present a diverse collection of examples where randomness plays a role.

1.1 Biometry: iris recognition

Biometry is the art of identifying a person on the basis of his or her personal biological characteristics, such as fingerprints or voice. From recent research it appears that with the human iris one can beat all existing automatic human identification systems. Iris recognition technology is based on the visible qualities of the iris. It converts these—via a video camera—into an "iris code" consisting of just 2048 bits. This is done in such a way that the code is hardly sensitive to the size of the iris or the size of the pupil. However, at different times and different places the iris code of the same person will not be exactly the same. Thus one has to allow for a certain percentage of mismatching bits when identifying a person. In fact, the system allows about 34% mismatches! How can this lead to a reliable identification system? The miracle is that different persons have very different irides. In particular, over a large collection of different irides the code bits take the values 0 and 1 about half of the time. But that is certainly not sufficient: if one bit would determine the other 2047, then we could only distinguish two persons. In other words, single bits may be random, but the correlation between bits is also crucial (we will discuss correlation at length in Chapter 10). John Daugman who has developed the iris recognition technology made comparisons between 222 743 pairs of iris

codes and concluded that of the 2048 bits 266 may be considered as uncorrelated ([6]). He then argues that we may consider an iris code as the result of 266 coin tosses with a fair coin. This implies that if we compare two such codes from different persons, then there is an astronomically small probability that these two differ in less than 34% of the bits—almost all pairs will differ in about 50% of the bits. This is illustrated in Figure 1.1, which originates from [6], and was kindly provided by John Daugman. The iris code data consist of numbers between 0 and 1, each a Hamming distance (the fraction of mismatches) between two iris codes. The data have been summarized in two histograms, that is, two graphs that show the number of counts of Hamming distances falling in a certain interval. We will encounter histograms and other summaries of data in Chapter 15. One sees from the figure that for codes from the same iris (left side) the mismatch fraction is only about 0.09, while for different irides (right side) it is about 0.46.

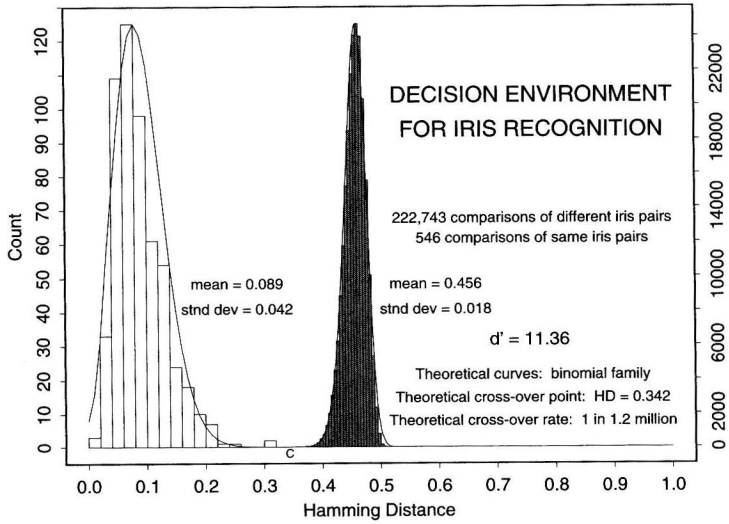


Fig. 1.1. Comparison of same and different iris pairs.

Source: J.Daugman. *Second IMA Conference on Image Processing: Mathematical Methods, Algorithms and Applications*, 2000. © Ellis Horwood Publishing Limited.

You may still wonder how it is possible that irides distinguish people so well. What about twins, for instance? The surprising thing is that although the color of eyes is hereditary, many features of iris patterns seem to be produced by so-called epigenetic events. This means that during embryo development the iris structure develops randomly. In particular, the iris patterns of (monozygotic) twins are as discrepant as those of two arbitrary individuals.