Springer Texts in Statistics

Advisors:
George Casella Stephen Fienberg Ingram Olkin

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

Whilst we have made considerable efforts to contact all holders of copyright material contained in this book, we may have failed to locate some of them. Should holders wish to contact the Publisher, we will be happy to come to some arrangement with them.

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

Apart from any fair dealing for the purposes of research or private study, or criticism or review, as permitted under the Copyright, Designs and Patents Act 1988, this publication may only be reproduced, stored or transmitted, in any form or by any means, with the prior permission in writing of the publishers, or in the case of reprographic reproduction in accordance with the terms of licences issued by the Copyright Licensing Agency. Enquiries concerning reproduction outside those terms should be sent to the publishers.

ISBN 978-1-85233-896-1

Springer Science+Business Media
springeronline.com

© Springer-Verlag London Limited 2005

The use of registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant laws and regulations and therefore free for general use.

The publisher makes no representation, express or implied, with regard to the accuracy of the information contained in this book and cannot accept any legal responsibility or liability for any errors or omissions that may be made.

12/3830/543210 Printed on acid-free paper SPIN 10943403

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.

Much of the material in this book would benefit from illustration with a computer using statistical software. A complete course should also involve computer exercises. Topics like simulation, the law of large numbers, the central limit theorem, and the bootstrap loudly call for this kind of experience. For this purpose, all the datasets discussed in the book are available at http://www.springeronline.com/1-85233-896-2. The same Web site also provides access, for instructors, to a complete set of solutions to the exercises; go to the Springer online catalog or contact textbooks@springer-sbm.com to apply for your password.

Delft, The Netherlands
January 2005

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

Contents

1 Why probability and statistics? 1
- 1.1 Biometry: iris recognition 1
- 1.2 Killer football 3
- 1.3 Cars and goats: the Monty Hall dilemma 4
- 1.4 The space shuttle *Challenger* 5
- 1.5 Statistics versus intelligence agencies 7
- 1.6 The speed of light 9

2 Outcomes, events, and probability 13
- 2.1 Sample spaces 13
- 2.2 Events 14
- 2.3 Probability 16
- 2.4 Products of sample spaces 18
- 2.5 An infinite sample space 19
- 2.6 Solutions to the quick exercises 21
- 2.7 Exercises 21

3 Conditional probability and independence 25
- 3.1 Conditional probability 25
- 3.2 The multiplication rule 27
- 3.3 The law of total probability and Bayes' rule 30
- 3.4 Independence 32
- 3.5 Solutions to the quick exercises 35
- 3.6 Exercises 37

4 Discrete random variables 41
4.1 Random variables 41
4.2 The probability distribution of a discrete random variable 43
4.3 The Bernoulli and binomial distributions 45
4.4 The geometric distribution 48
4.5 Solutions to the quick exercises 50
4.6 Exercises 51

5 Continuous random variables 57
5.1 Probability density functions 57
5.2 The uniform distribution 60
5.3 The exponential distribution 61
5.4 The Pareto distribution 63
5.5 The normal distribution 64
5.6 Quantiles 65
5.7 Solutions to the quick exercises 67
5.8 Exercises 68

6 Simulation 71
6.1 What is simulation? 71
6.2 Generating realizations of random variables 72
6.3 Comparing two jury rules 75
6.4 The single-server queue 80
6.5 Solutions to the quick exercises 84
6.6 Exercises 85

7 Expectation and variance 89
7.1 Expected values 89
7.2 Three examples 93
7.3 The change-of-variable formula 94
7.4 Variance 96
7.5 Solutions to the quick exercises 99
7.6 Exercises 99

8 Computations with random variables 103
8.1 Transforming discrete random variables 103
8.2 Transforming continuous random variables 104
8.3 Jensen's inequality 106

8.4 Extremes . . . 108
8.5 Solutions to the quick exercises . . . 110
8.6 Exercises . . . 111

9 Joint distributions and independence . . . 115
9.1 Joint distributions of discrete random variables . . . 115
9.2 Joint distributions of continuous random variables . . . 118
9.3 More than two random variables . . . 122
9.4 Independent random variables . . . 124
9.5 Propagation of independence . . . 125
9.6 Solutions to the quick exercises . . . 126
9.7 Exercises . . . 127

10 Covariance and correlation . . . 135
10.1 Expectation and joint distributions . . . 135
10.2 Covariance . . . 138
10.3 The correlation coefficient . . . 141
10.4 Solutions to the quick exercises . . . 143
10.5 Exercises . . . 144

11 More computations with more random variables . . . 151
11.1 Sums of discrete random variables . . . 151
11.2 Sums of continuous random variables . . . 154
11.3 Product and quotient of two random variables . . . 159
11.4 Solutions to the quick exercises . . . 162
11.5 Exercises . . . 163

12 The Poisson process . . . 167
12.1 Random points . . . 167
12.2 Taking a closer look at random arrivals . . . 168
12.3 The one-dimensional Poisson process . . . 171
12.4 Higher-dimensional Poisson processes . . . 173
12.5 Solutions to the quick exercises . . . 176
12.6 Exercises . . . 176

13 The law of large numbers . . . 181
13.1 Averages vary less . . . 181
13.2 Chebyshev's inequality . . . 183

13.3 The law of large numbers . . . 185
13.4 Consequences of the law of large numbers . . . 188
13.5 Solutions to the quick exercises . . . 191
13.6 Exercises . . . 191

14 The central limit theorem . . . 195
14.1 Standardizing averages . . . 195
14.2 Applications of the central limit theorem . . . 199
14.3 Solutions to the quick exercises . . . 202
14.4 Exercises . . . 203

15 Exploratory data analysis: graphical summaries . . . 207
15.1 Example: the Old Faithful data . . . 207
15.2 Histograms . . . 209
15.3 Kernel density estimates . . . 212
15.4 The empirical distribution function . . . 219
15.5 Scatterplot . . . 221
15.6 Solutions to the quick exercises . . . 225
15.7 Exercises . . . 226

16 Exploratory data analysis: numerical summaries . . . 231
16.1 The center of a dataset . . . 231
16.2 The amount of variability of a dataset . . . 233
16.3 Empirical quantiles, quartiles, and the IQR . . . 234
16.4 The box-and-whisker plot . . . 236
16.5 Solutions to the quick exercises . . . 238
16.6 Exercises . . . 240

17 Basic statistical models . . . 245
17.1 Random samples and statistical models . . . 245
17.2 Distribution features and sample statistics . . . 248
17.3 Estimating features of the "true" distribution . . . 253
17.4 The linear regression model . . . 256
17.5 Solutions to the quick exercises . . . 259
17.6 Exercises . . . 259

18 The bootstrap 269
18.1 The bootstrap principle 269
18.2 The empirical bootstrap 272
18.3 The parametric bootstrap 276
18.4 Solutions to the quick exercises 279
18.5 Exercises 280

19 Unbiased estimators 285
19.1 Estimators 285
19.2 Investigating the behavior of an estimator 287
19.3 The sampling distribution and unbiasedness 288
19.4 Unbiased estimators for expectation and variance 292
19.5 Solutions to the quick exercises 294
19.6 Exercises 294

20 Efficiency and mean squared error 299
20.1 Estimating the number of German tanks 299
20.2 Variance of an estimator 302
20.3 Mean squared error 305
20.4 Solutions to the quick exercises 307
20.5 Exercises 307

21 Maximum likelihood 313
21.1 Why a general principle? 313
21.2 The maximum likelihood principle 314
21.3 Likelihood and loglikelihood 316
21.4 Properties of maximum likelihood estimators 321
21.5 Solutions to the quick exercises 322
21.6 Exercises 323

22 The method of least squares 329
22.1 Least squares estimation and regression 329
22.2 Residuals 332
22.3 Relation with maximum likelihood 335
22.4 Solutions to the quick exercises 336
22.5 Exercises 337

23 Confidence intervals for the mean 341
23.1 General principle 341
23.2 Normal data 345
23.3 Bootstrap confidence intervals 350
23.4 Large samples 353
23.5 Solutions to the quick exercises 355
23.6 Exercises 356

24 More on confidence intervals 361
24.1 The probability of success 361
24.2 Is there a general method? 364
24.3 One-sided confidence intervals 366
24.4 Determining the sample size 367
24.5 Solutions to the quick exercises 368
24.6 Exercises 369

25 Testing hypotheses: essentials 373
25.1 Null hypothesis and test statistic 373
25.2 Tail probabilities 376
25.3 Type I and type II errors 377
25.4 Solutions to the quick exercises 379
25.5 Exercises 380

26 Testing hypotheses: elaboration 383
26.1 Significance level 383
26.2 Critical region and critical values 386
26.3 Type II error 390
26.4 Relation with confidence intervals 392
26.5 Solutions to the quick exercises 393
26.6 Exercises 394

27 The *t*-test 399
27.1 Monitoring the production of ball bearings 399
27.2 The one-sample *t*-test 401
27.3 The *t*-test in a regression setting 405
27.4 Solutions to the quick exercises 409
27.5 Exercises 410

28 Comparing two samples 415
28.1 Is dry drilling faster than wet drilling? 415
28.2 Two samples with equal variances 416
28.3 Two samples with unequal variances 419
28.4 Large samples 422
28.5 Solutions to the quick exercises 424
28.6 Exercises 424

A Summary of distributions 429

B Tables of the normal and *t*-distributions 431

C Answers to selected exercises 435

D Full solutions to selected exercises 445

References 475

List of symbols 477

Index 479