

UNITEXT 139



Alberto Rotondi · Paolo Pedroni
Antonio Pievatolo

Probability, Statistics and Simulation

With Application Programs
Written in R

 Springer

UNITEXT

La Matematica per il 3+2

Volume 139

Editor-in-Chief

Alfio Quarteroni, Politecnico di Milano, Milan, Italy; École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

Series Editors

Luigi Ambrosio, Scuola Normale Superiore, Pisa, Italy

Paolo Biscari, Politecnico di Milano, Milan, Italy

Ciro Ciliberto, Università di Roma “Tor Vergata”, Rome, Italy

Camillo De Lellis, Institute for Advanced Study, Princeton, New Jersey, USA

Massimiliano Gubinelli, Hausdorff Center for Mathematics, Rheinische Friedrich-Wilhelms-Universität, Bonn, Germany

Victor Panaretos, Institute of Mathematics, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

Lorenzo Rosasco, DIBRIS, Università degli Studi di Genova, Genova, Italy; Center for Brains Mind and Machines, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA; Istituto Italiano di Tecnologia, Genova, Italy

The **UNITEXT - La Matematica per il 3+2** series is designed for undergraduate and graduate academic courses, and also includes advanced textbooks at a research level.

Originally released in Italian, the series now publishes textbooks in English addressed to students in mathematics worldwide.

Some of the most successful books in the series have evolved through several editions, adapting to the evolution of teaching curricula.

Submissions must include at least 3 sample chapters, a table of contents, and a preface outlining the aims and scope of the book, how the book fits in with the current literature, and which courses the book is suitable for.

For any further information, please contact the Editor at Springer: francesca.bonadei@springer.com

THE SERIES IS INDEXED IN SCOPUS

UNITEXT is glad to announce a new series of **free webinars and interviews** handled by the Board members, who will rotate in order to interview top experts in their field.

In the first session, going live on June 9, Alfio Quarteroni will interview Luigi Ambrosio. The speakers will dive into the subject of Optimal Transport, and will discuss the most challenging open problems and the future developments in the field.

Click here to subscribe to the event!

<https://cassyni.com/events/TPQ2UgkCbJvvz5QbkcWXo3>

Alberto Rotondi • Paolo Pedroni •
Antonio Pievatolo

Probability, Statistics and Simulation

With Application Programs Written in R

 Springer

Alberto Rotondi
Dipartimento di Fisica
Università di Pavia
Pavia, Italy

Paolo Pedroni
Istituto Nazionale di Fisica Nucleare
Università di Pavia
Pavia, Italy

Antonio Pievatolo
Istituto di Matematica Applicata e
Tecnologie Informatiche
Consiglio Nazionale delle Ricerche
Milano, Italy

ISSN 2038-5714

UNITEXT

ISSN 2038-5722

La Matematica per il 3+2

ISBN 978-3-031-09428-6

<https://doi.org/10.1007/978-3-031-09429-3>

ISSN 2532-3318 (electronic)

ISSN 2038-5757 (electronic)

ISBN 978-3-031-09429-3 (eBook)

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

This work is subject to copyright. All rights are solely and exclusively licensed 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, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Cover illustration: “The face number three and two numbers three, one chance over 1326” (photo by the authors)

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This book, based on the fourth Italian edition, comes from the collaboration between two experimental physicists and one statistician. Among non-statisticians, physicists are perhaps the ones who most appreciate and use probability and statistics, but most of the time in a pragmatic and manual way, having in mind the solution of specific problems or technical applications. On the other hand, in the crucial comparison between theory and experiment, it is sometimes necessary to use sophisticated methods which require knowledge of the fundamental logical and mathematical principles at the basis of the study of random phenomena. More generally, even those who are not statisticians have often to face, in any research field, problems that require particular attention and expertise for the treatment of random or aleatory aspects. These skills are naturally mastered by the statistician, whose research interests are the laws of chance.

This text has been prepared with the aim to seek a synthesis between these different approaches, to provide the reader not only with tools useful to address problems, but also with a guide to the correct methodologies needed to understand the complicated and fascinating world of random phenomena.

Such an objective obviously involved choices, sometimes even painful, both of content type and style. As for style, we tried not to give up the precision needed to properly teach the important concepts. When treating applications, we privileged the methods that do not require excessive preliminary conceptual elaborations.

As an example, we have tried to use, whenever possible, approximate methods for interval estimation, with Gaussian approximations for the estimator distributions. Similarly, in the case of least squares, we have extensively adopted the approximation based on the χ^2 distribution to verify the fitting of a model to the data.

We also avoided insisting on the formal treatment of complicated problems in cases where a solution using computer and simple simulation programs could be easily found.

In our book, simulation plays an important role in the presentation of many topics and in the verification of the accuracy of many techniques and approximations. This

feature, already present in the first Italian edition, is now common to many data science texts and, in our opinion, confirms the validity of our initial choice.

This book is aimed primarily at students of scientific undergraduate courses, such as engineering, computer science, and physics. However, we think that it can also be useful to all those scientific researchers who have to solve practical problems involving probabilistic, statistical, and simulation aspects. For this reason, we have given space to some topics, such as Monte Carlo methods and their applications, minimization techniques, and data analysis methods, which, usually, are only briefly mentioned in introductory texts.

The mathematical knowledge required by the reader is that which is normally given in the teaching of the basic calculus course in the scientific degrees, with the addition of minimum notions of linear algebra and advanced calculus, such as the elementary concepts of the derivation and integration of multidimensional functions.

The structure of the text allows different learning paths and reading levels. The first seven chapters deal with all the topics usually developed in a standard, basic statistical course. At the choice of the teacher, this program can be integrated with some more advanced topics from the other chapters. For example, Chap. 8 should certainly be included in a simulation-oriented course.

The notions of probability and statistics usually taught to physics students in undergraduate laboratory courses are enclosed in the first three chapters, in Chaps. 6 and 7 (basic statistics) and in Chap. 12, written explicitly for physicists and for all those who need to process data from laboratory experiments.

Many pages are devoted to the complete resolution of several exercises inserted directly inside the chapters to better explain the covered topics. We also recommend to the reader the problems (all with solutions) reported at the end of each chapter.

This book makes use of the statistical software R, which has now become the world standard for solving statistical problems. The 2019 ranking of the Institute of American Electrical and Electronic Engineers (IEEE) places R in fourth position among the most popular programming languages, after Python, Java, and C. Many R routines have been written by us, to guide the reader while going through the text. These routines can be easily downloaded from the link specified below. We therefore recommend an interactive reading, in which the study of a topic is followed by the use of R routines in the way showed both in the text and in the technical instructions included in the indicated Web pages.

We thank again the readers who reported errors or inaccuracies present in the previous Italian editions, and the publisher, Springer, for the continued trust in our work.

Pavia, Italy
Pavia, Italy
Milano, Italy
March 2022

Alberto Rotondi
Paolo Pedroni
Antonio Pievatolo

How to Use the Text

Figures, equations, definitions, theorems, Tables, and exercises are numbered progressively.

The abbreviations of quotations (e.g., [57]) refer to the bibliographic list at the end of the book.

Solutions of the problems are given in Appendix D. The table of symbols reported in Appendix A may also be useful.

Calculation codes as `hist` are marked with a different text style. Routines starting with a lowercase letter are (with some exceptions) the original R codes, which can be freely copied from the CRAN (Comprehensive R Archive Network) website, while those starting with an uppercase letter are written by the authors and are in:

```
https://tinyurl.com/ProbStatSimul
```

In this site, you will also find all the information for the installation and the use of R, a guide to the use of routines written by the authors and complementary teaching materials.

Contents

1	Probability	1
1.1	Chance, Chaos and Determinism	1
1.2	Some Basic Terms	8
1.3	The Concept of Probability	10
1.4	Axiomatic Probability	13
1.5	Repeated Trials	19
1.6	Elements of Combinatorial Analysis	23
1.7	Bayes' Theorem	26
1.8	Learning Algorithms	33
1.9	Problems	36
2	Representation of Random Phenomena	39
2.1	Introduction	39
2.2	Random Variables	40
2.3	Cumulative or Distribution Function	44
2.4	Data Representation	46
2.5	Discrete Random Variables	49
2.6	Binomial Distribution	51
2.7	Continuous Random Variables	54
2.8	Mean, Sum of Squares, Variance, Standard Deviation and Quantiles	57
2.9	Operators	64
2.10	Simple Random Sample	67
2.11	Convergence Criteria	69
2.12	Problems	73
3	Basic Probability Theory	75
3.1	Introduction	75
3.2	Properties of the Binomial Distribution	75
3.3	Poisson Distribution	79
3.4	Normal or Gaussian Density	81

3.5	The Three-Sigma Law and the Standard Gaussian Density	87
3.6	Central Limit Theorem and Universality of the Gaussian Curve	91
3.7	Poisson Stochastic Processes	94
3.8	χ^2 Density	101
3.9	Uniform Density	108
3.10	Chebyshev's Inequality	112
3.11	How to Use Probability Calculus	113
3.12	Problems	122
4	Multivariate Probability Theory	125
4.1	Introduction	125
4.2	Multivariate Statistical Distributions	126
4.3	Covariance and Correlation	134
4.4	Two-Dimensional Gaussian Distribution	139
4.5	The General Multidimensional Case	148
4.6	Multivariate Probability Regions	155
4.7	Multinomial Distribution	159
4.8	Problems	161
5	Functions of Random Variables	163
5.1	Introduction	163
5.2	Functions of a Random Variable	165
5.3	Functions of Several Random Variables	168
5.4	Mean and Variance Transformation	184
5.5	Means and Variances for n Variables	190
5.6	Problems	197
6	Basic Statistics: Parameter Estimation	199
6.1	Introduction	199
6.2	Confidence Intervals	201
6.3	Confidence Intervals with Pivotal Variables	204
6.4	Mention of the Bayesian Approach	207
6.5	Some Notations	208
6.6	Probability Estimation	209
6.7	Probability Estimation from Large Samples	212
6.8	Poissonian Interval Estimation	218
6.9	Mean Estimation from Large Samples	222
6.10	Variance Estimation from Large Samples	224
6.11	Mean and Variance Estimation for Gaussian Samples	229
6.12	How to Use the Estimation Theory	232
6.13	Estimates from a Finite Population	239
6.14	Histogram Analysis	242
6.15	Estimation of the Correlation	248
6.16	Problems	257

7	Basic Statistics: Hypothesis Testing	259
7.1	Testing One Hypothesis	259
7.2	The Gaussian z -Test	263
7.3	Student's t -Test	269
7.4	Chi-Square Test	274
7.5	Compatibility Check Between Sample and Population	277
7.6	Hypothesis Testing with Contingency Tables	285
7.7	Multiple Tests	291
7.8	Snedecor's F -Test	298
7.9	Analysis of Variance (ANOVA)	299
7.10	Two-Way ANOVA	309
7.11	Problems	315
8	Monte Carlo Methods	319
8.1	Introduction	319
8.2	What Is Monte Carlo?	320
8.3	Mathematical Aspects	323
8.4	Generation of Discrete Random Variables	324
8.5	Generation of Continuous Random Variables	328
8.6	Linear Search Method	334
8.7	Rejection Method	336
8.8	Particular Random Generation Methods	343
8.9	Monte Carlo Analysis of Distributions	348
8.10	Evaluation of Confidence Intervals	351
8.11	Simulation of Counting Experiments	355
8.12	Non-parametric Bootstrap	359
8.13	Hypothesis Test with Simulated Data	364
8.14	Problems	366
9	Applications of Monte Carlo Methods	369
9.1	Introduction	369
9.2	Study of Diffusion Phenomena	369
9.3	Simulation of Stochastic Processes	377
9.4	Number of Workers in a Plant: Synchronous Simulation	382
9.5	Number of Workers in a Plant: Asynchronous Simulation	385
9.6	Kolmogorov-Smirnov Test	388
9.7	Metropolis Algorithm	393
9.8	Ising Model	397
9.9	Definite Integral Calculation	400
9.10	Importance Sampling	404
9.11	Stratified Sampling	405
9.12	Multidimensional Integrals	410
9.13	Problems	410

10	Statistical Inference and Likelihood	413
10.1	Introduction	413
10.2	Maximum Likelihood (ML) Method	415
10.3	Estimator Properties	420
10.4	Theorems on Estimators	423
10.5	Confidence Intervals	434
10.6	Least Squares Method and Maximum Likelihood	437
10.7	Best Fit of Densities to Data and Histograms	439
10.8	Weighted Mean	444
10.9	Test of Hypotheses	450
10.10	One- or Two-Sample Tests	452
10.11	Most Powerful Tests	456
10.12	Test Functions	459
10.13	Sequential Tests	465
10.14	Problems	471
11	Least Squares	475
11.1	Introduction	475
11.2	No Errors on Predictors	477
11.3	Errors in Predictors	481
11.4	Least Squares Regression Lines: Unweighted Case	484
11.5	Unweighted Linear Least Squares	491
11.6	Weighted Linear Least Squares	495
11.7	Properties of Least Squares Estimates	499
11.8	Model Testing and Search for Functional Forms	502
11.9	Search for Correlations	511
11.10	Fit Strategies	516
11.11	Nonlinear Least Squares	517
11.12	Problems	520
12	Experimental Data Analysis	523
12.1	Introduction	523
12.2	Terminology	524
12.3	Constant and Variable Physical Quantities	525
12.4	Instrumental Sensitivity and Accuracy	526
12.5	Measurement Uncertainty	529
12.6	Treatment of Systematic Effects	532
12.7	Best Fit with Offset Systematic Errors	536
12.8	Best Fit with Scale Systematic Errors	540
12.9	Indirect Measurements and Error Propagation	542
12.10	Measurement Types	551
12.11	$M(\mathbf{0}, \mathbf{0}, \Delta)$ Measurements	552
12.12	$M(\mathbf{0}, \sigma, \mathbf{0})$ Measurements	553
12.13	$M(\mathbf{0}, \sigma, \Delta)$ Measurements	556
12.14	$M(f, \mathbf{0}, \mathbf{0})$ Measurements	558

- 12.15 $M(f, \sigma, \mathbf{0}), M(f, \mathbf{0}, \Delta)$ and $M(f, \sigma, \Delta)$
Measurements 565
- 12.16 A Case Study: Millikan's Experiments 569
- 12.17 Some Remarks on the Scientific Method 572
- 12.18 Problems 578

- A Table of Symbols** 581
- B R Software** 583
- C Moment-Generating Functions** 587
- D Solutions of Problems** 591

- E Tables** 615
 - E.1 Integral of the Gaussian Density 615
 - E.2 Quantiles of the Student's Density 616
 - E.3 Integrals of the Reduced χ^2 Density 616
 - E.4 Quantile Values of the Non-Reduced χ^2 Density 616
 - E.5 Quantiles of the F Density 617

- Bibliography** 625
- Index** 631