

# Lecture Notes in Physics

Volume 941

## *Founding Editors*

W. Beiglböck  
J. Ehlers  
K. Hepp  
H. Weidenmüller

## *Editorial Board*

M. Bartelmann, Heidelberg, Germany  
P. Hänggi, Augsburg, Germany  
M. Hjorth-Jensen, Oslo, Norway  
R.A.L. Jones, Sheffield, UK  
M. Lewenstein, Barcelona, Spain  
H. von Löhneysen, Karlsruhe, Germany  
A. Rubio, Hamburg, Germany  
M. Salmhofer, Heidelberg, Germany  
W. Schleich, Ulm, Germany  
S. Theisen, Potsdam, Germany  
D. Vollhardt, Augsburg, Germany  
J.D. Wells, Ann Arbor, USA  
G.P. Zank, Huntsville, USA

# The Lecture Notes in Physics

The series Lecture Notes in Physics (LNP), founded in 1969, reports new developments in physics research and teaching—quickly and informally, but with a high quality and the explicit aim to summarize and communicate current knowledge in an accessible way. Books published in this series are conceived as bridging material between advanced graduate textbooks and the forefront of research and to serve three purposes:

- to be a compact and modern up-to-date source of reference on a well-defined topic
- to serve as an accessible introduction to the field to postgraduate students and nonspecialist researchers from related areas
- to be a source of advanced teaching material for specialized seminars, courses and schools

Both monographs and multi-author volumes will be considered for publication. Edited volumes should, however, consist of a very limited number of contributions only. Proceedings will not be considered for LNP.

Volumes published in LNP are disseminated both in print and in electronic formats, the electronic archive being available at [springerlink.com](http://springerlink.com). The series content is indexed, abstracted and referenced by many abstracting and information services, bibliographic networks, subscription agencies, library networks, and consortia.

Proposals should be sent to a member of the Editorial Board, or directly to the managing editor at Springer:

Christian Caron  
Springer Heidelberg  
Physics Editorial Department I  
Tiergartenstrasse 17  
69121 Heidelberg/Germany  
[christian.caron@springer.com](mailto:christian.caron@springer.com)

More information about this series at <http://www.springer.com/series/5304>

Luca Lista

# Statistical Methods for Data Analysis in Particle Physics

Second Edition

 Springer

Luca Lista  
INFN Sezione di Napoli  
Napoli, Italy

ISSN 0075-8450

ISSN 1616-6361 (electronic)

Lecture Notes in Physics

ISBN 978-3-319-62839-4

ISBN 978-3-319-62840-0 (eBook)

DOI 10.1007/978-3-319-62840-0

Library of Congress Control Number: 2017948232

© Springer International Publishing AG 2016, 2017

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. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by Springer Nature

The registered company is Springer International Publishing AG

The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

# Preface

This book started as a collection of material from a course of lectures on Statistical Methods for Data Analysis I gave to Ph.D. students in physics at the University of Naples Federico II from 2009 to 2017 and was subsequently enriched with material from other seminars and lectures I have been invited to give in the last years.

The aim of the book is to present and elaborate the main concepts and tools that physicists use to analyze experimental data.

An introduction to probability theory and basic statistics is provided mainly as refresher lectures to students who did not take a formal course on statistics before starting their Ph.D. This also gives the opportunity to introduce Bayesian approach to probability, which is a new topic to many students.

More advanced topics follow, up to recent developments in statistical methods used for particle physics, in particular for data analyses at the Large Hadron Collider.

Many of the covered tools and methods have applications in high-energy physics, but their scope could well be extended to other fields.

A shorter version of the course was presented at CERN in November 2009 as lectures on Statistical Methods in LHC Data Analysis for the ATLAS and CMS experiments. The chapter that discusses discoveries and upper limits was improved after the lectures on the subject I gave in Autrans, France, at the IN2P3 School of Statistics in May 2012. I was also invited to conduct a seminar about Statistical Methods at Gent University, Belgium, in October 2014, which gave me the opportunity to review some of my material and add new examples.

## Note to the Second Edition

The second edition of this book reflects the work I did in preparation of the lectures that I was invited to give during the CERN-JINR European School of High-Energy Physics (15–28 June 2016, Skeikampen, Norway). On that occasion, I reviewed, expanded, and reordered my material.

In addition, with respect to the first edition, I added a chapter about unfolding, an extended discussion about the best linear unbiased estimator, and an introduction to machine learning algorithms, in particular artificial neural networks, with hints about deep learning, and boosted decision trees.

## **Acknowledgments**

I am grateful to Louis Lyons who carefully and patiently read the first edition of my book and provided useful comments and suggestions. I would like to thank Eliam Gross for providing useful examples and for reviewing the sections about the look elsewhere effect. I also received useful comments from Vitaliano Ciulli and from Luis Isaac Ramos Garcia.

I considered all feedback I received in the preparation of this second edition.

Napoli, Italy

Luca Lista

# Contents

<b>1</b>	<b>Probability Theory</b> .....	1
1.1	Why Probability Matters to a Physicist .....	1
1.2	The Concept of Probability .....	2
1.3	Repeatable and Non-Repeatable Cases .....	2
1.4	Different Approaches to Probability .....	3
1.5	Classical Probability .....	4
1.6	Generalization to the Continuum .....	6
1.6.1	The Bertrand's Paradox .....	7
1.7	Axiomatic Probability Definition .....	8
1.8	Probability Distributions .....	9
1.9	Conditional Probability .....	9
1.10	Independent Events .....	10
1.11	Law of Total Probability .....	11
1.12	Average, Variance and Covariance .....	12
1.13	Transformations of Variables .....	15
1.14	The Bernoulli Process .....	16
1.15	The Binomial Process .....	17
1.16	Multinomial Distribution .....	20
1.17	The Law of Large Numbers .....	21
1.18	Frequentist Definition of Probability .....	22
	References .....	23
<b>2</b>	<b>Probability Distribution Functions</b> .....	25
2.1	Introduction .....	25
2.2	Definition of Probability Distribution Function .....	25
2.3	Average and Variance in the Continuous Case .....	27
2.4	Mode, Median, Quantiles .....	28
2.5	Cumulative Distribution .....	28
2.6	Continuous Transformations of Variables .....	29
2.7	Uniform Distribution .....	30

2.8	Gaussian Distribution .....	31
2.9	$\chi^2$ Distribution .....	32
2.10	Log Normal Distribution .....	33
2.11	Exponential Distribution.....	34
2.12	Poisson Distribution .....	35
2.13	Other Distributions Useful in Physics .....	41
	2.13.1 Breit–Wigner Distribution .....	41
	2.13.2 Relativistic Breit–Wigner Distribution.....	42
	2.13.3 Argus Function.....	43
	2.13.4 Crystal Ball Function .....	44
	2.13.5 Landau Distribution.....	46
2.14	Central Limit Theorem .....	46
2.15	Probability Distribution Functions in More than One Dimension .....	49
	2.15.1 Marginal Distributions.....	49
	2.15.2 Independent Variables .....	50
	2.15.3 Conditional Distributions.....	53
2.16	Gaussian Distributions in Two or More Dimensions .....	54
	References.....	58
<b>3</b>	<b>Bayesian Approach to Probability .....</b>	<b>59</b>
3.1	Introduction .....	59
3.2	Bayes’ Theorem.....	59
3.3	Bayesian Probability Definition .....	64
3.4	Bayesian Probability and Likelihood Functions.....	67
	3.4.1 Repeated Use of Bayes’ Theorem and Learning Process .....	67
3.5	Bayesian Inference.....	68
	3.5.1 Parameters of Interest and Nuisance Parameters .....	69
	3.5.2 Credible Intervals .....	70
3.6	Bayes Factors .....	73
3.7	Subjectiveness and Prior Choice .....	74
3.8	Jeffreys’ Prior .....	75
3.9	Reference Priors .....	76
3.10	Improper Priors .....	76
3.11	Transformations of Variables and Error Propagation .....	79
	References.....	79
<b>4</b>	<b>Random Numbers and Monte Carlo Methods .....</b>	<b>81</b>
4.1	Pseudorandom Numbers.....	81
4.2	Pseudorandom Generators Properties.....	82
4.3	Uniform Random Number Generators .....	84
	4.3.1 Remapping Uniform Random Numbers .....	85
4.4	Discrete Random Number Generators .....	85



4.5	Nonuniform Random Number Generators.....	86
4.5.1	Nonuniform Distribution from Inversion of the Cumulative Distribution .....	86
4.5.2	Gaussian Generator Using the Central Limit Theorem .....	88
4.5.3	Gaussian Generator with the Box–Muller Method.....	89
4.6	Monte Carlo Sampling.....	89
4.6.1	Hit-or-Miss Monte Carlo .....	90
4.6.2	Importance Sampling .....	91
4.7	Numerical Integration with Monte Carlo Methods.....	92
4.8	Markov Chain Monte Carlo .....	93
	References.....	95
<b>5</b>	<b>Parameter Estimate .....</b>	<b>97</b>
5.1	Introduction .....	97
5.2	Inference.....	97
5.3	Parameters of Interest .....	98
5.4	Nuisance Parameters .....	98
5.5	Measurements and Their Uncertainties .....	99
5.5.1	Statistical and Systematic Uncertainties .....	99
5.6	Frequentist vs Bayesian Inference .....	100
5.7	Estimators .....	100
5.8	Properties of Estimators .....	101
5.8.1	Consistency .....	102
5.8.2	Bias .....	102
5.8.3	Minimum Variance Bound and Efficiency .....	102
5.8.4	Robust Estimators.....	103
5.9	Binomial Distribution for Efficiency Estimate .....	104
5.10	Maximum Likelihood Method .....	105
5.10.1	Likelihood Function .....	105
5.10.2	Extended Likelihood Function .....	106
5.10.3	Gaussian Likelihood Functions .....	108
5.11	Errors with the Maximum Likelihood Method .....	109
5.11.1	Second Derivatives Matrix .....	109
5.11.2	Likelihood Scan .....	110
5.11.3	Properties of Maximum Likelihood Estimators .....	112
5.12	Minimum $\chi^2$ and Least-Squares Methods .....	114
5.12.1	Linear Regression .....	115
5.12.2	Goodness of Fit and $p$ -Value .....	118
5.13	Binned Data Samples .....	118
5.13.1	Minimum $\chi^2$ Method for Binned Histograms .....	119
5.13.2	Binned Poissonian Fits .....	120

5.14	Error Propagation .....	121
5.14.1	Simple Cases of Error Propagation .....	121
5.15	Treatment of Asymmetric Errors .....	123
5.15.1	Asymmetric Error Combination with a Linear Model .....	124
	References .....	127
<b>6</b>	<b>Combining Measurements</b> .....	<b>129</b>
6.1	Introduction .....	129
6.2	Simultaneous Fits and Control Regions .....	129
6.3	Weighted Average .....	131
6.4	$\chi^2$ in $n$ Dimensions .....	132
6.5	The Best Linear Unbiased Estimator .....	133
6.5.1	Quantifying the Importance of Individual Measurements .....	135
6.5.2	Negative Weights .....	137
6.5.3	Iterative Application of the BLUE Method .....	139
	References .....	140
<b>7</b>	<b>Confidence Intervals</b> .....	<b>143</b>
7.1	Introduction .....	143
7.2	Neyman Confidence Intervals .....	143
7.2.1	Construction of the Confidence Belt .....	144
7.2.2	Inversion of the Confidence Belt .....	146
7.3	Binomial Intervals .....	147
7.4	The Flip-Flopping Problem .....	150
7.5	The Unified Feldman–Cousins Approach .....	152
	References .....	154
<b>8</b>	<b>Convolution and Unfolding</b> .....	<b>155</b>
8.1	Introduction .....	155
8.2	Convolution .....	155
8.2.1	Convolution and Fourier Transform .....	156
8.2.2	Discrete Convolution and Response Matrix .....	158
8.2.3	Efficiency and Background .....	158
8.3	Unfolding by Inversion of the Response Matrix .....	160
8.4	Bin-by-Bin Correction Factors .....	163
8.5	Regularized Unfolding .....	163
8.5.1	Tikhonov Regularization .....	164
8.6	Iterative Unfolding .....	166
8.6.1	Treatment of Background .....	171
8.7	Other Unfolding Methods .....	171
8.8	Software Implementations .....	173
8.9	Unfolding in More Dimensions .....	173
	References .....	173

- 9 Hypothesis Tests** ..... 175
  - 9.1 Introduction ..... 175
  - 9.2 Test Statistic ..... 175
  - 9.3 Type I and Type II Errors ..... 177
  - 9.4 Fisher’s Linear Discriminant ..... 178
  - 9.5 The Neyman–Pearson Lemma ..... 181
  - 9.6 Projective Likelihood Ratio Discriminant ..... 181
  - 9.7 Kolmogorov–Smirnov Test ..... 182
  - 9.8 Wilks’ Theorem ..... 184
  - 9.9 Likelihood Ratio in the Search for a New Signal ..... 185
  - 9.10 Multivariate Discrimination with Machine Learning ..... 188
    - 9.10.1 Overtraining ..... 189
  - 9.11 Artificial Neural Networks ..... 190
    - 9.11.1 Deep Learning ..... 192
    - 9.11.2 Convolutional Neural Networks ..... 193
  - 9.12 Boosted Decision Trees ..... 196
  - 9.13 Multivariate Analysis Implementations ..... 199
  - References ..... 203
- 10 Discoveries and Upper Limits** ..... 205
  - 10.1 Searches for New Phenomena: Discovery and Upper Limits ..... 205
  - 10.2 Claiming a Discovery ..... 206
    - 10.2.1  $p$ -Values ..... 206
    - 10.2.2 Significance Level ..... 207
    - 10.2.3 Significance and Discovery ..... 208
    - 10.2.4 Significance for Poissonian Counting Experiments ..... 208
    - 10.2.5 Significance with Likelihood Ratio ..... 209
    - 10.2.6 Significance Evaluation with Toy Monte Carlo ..... 210
  - 10.3 Excluding a Signal Hypothesis ..... 211
  - 10.4 Combined Measurements and Likelihood Ratio ..... 211
  - 10.5 Definitions of Upper Limit ..... 211
  - 10.6 Bayesian Approach ..... 212
    - 10.6.1 Bayesian Upper Limits for Poissonian Counting ..... 212
    - 10.6.2 Limitations of the Bayesian Approach ..... 215
  - 10.7 Frequentist Upper Limits ..... 215
    - 10.7.1 Frequentist Upper Limits for Counting Experiments ..... 216
    - 10.7.2 Frequentist Limits in Case of Discrete Variables ..... 217
    - 10.7.3 Feldman–Cousins Unified Approach ..... 218
  - 10.8 Modified Frequentist Approach: The  $CL_s$  Method ..... 221
  - 10.9 Presenting Upper Limits: The Brazil Plot ..... 225
  - 10.10 Nuisance Parameters and Systematic Uncertainties ..... 226
    - 10.10.1 Nuisance Parameters with the Bayesian Approach ..... 226
    - 10.10.2 Hybrid Treatment of Nuisance Parameters ..... 227
    - 10.10.3 Event Counting Uncertainties ..... 227

- 10.11 Upper Limits Using the Profile Likelihood ..... 228
- 10.12 Variations of the Profile-Likelihood Test Statistic ..... 229
  - 10.12.1 Test Statistic for Positive Signal Strength ..... 230
  - 10.12.2 Test Statistic for Discovery ..... 230
  - 10.12.3 Test Statistic for Upper Limits ..... 230
  - 10.12.4 Higgs Test Statistic ..... 231
  - 10.12.5 Asymptotic Approximations ..... 231
  - 10.12.6 Asimov Datasets ..... 231
- 10.13 The Look Elsewhere Effect..... 242
  - 10.13.1 Trial Factors ..... 243
  - 10.13.2 Look Elsewhere Effect in More Dimensions ..... 246
- References..... 248
  
- Index..... 251**

# List of Tables

Table 1.1	Possible values of the sum of two dice rolls with all possible pair combinations and corresponding probability .....	5
Table 2.1	Probabilities corresponding to $Z\sigma$ one-dimensional intervals and two-dimensional contours for different values of $Z$ .....	58
Table 3.1	Assessing evidence with Bayes factors according to the scale proposed in [4].....	74
Table 3.2	Jeffreys’ priors corresponding to the parameters of some of the most frequently used PDFs.....	76
Table 6.1	Properties of different indicators of a measurement’s importance within a BLUE combination .....	137
Table 10.1	Significance expressed as ‘ $Z\sigma$ ’ and corresponding $p$ -value in a number of typical cases .....	208
Table 10.2	Upper limits in presence of negligible background evaluated under the Bayesian approach for different number of observed events $n$ .....	213
Table 10.3	Upper and lower limits in presence of negligible background ( $b = 0$ ) with the Feldman–Cousins approach .....	219

# List of Examples

Example 1.1	Two Dice Roll Probability .....	5
Example 1.2	Combination of Detector Efficiencies .....	10
Example 1.3	Application to the Sum of Dice Rolls .....	16
Example 2.4	Strip Detectors .....	31
Example 2.5	Poisson Distributions as Limit of Binomial Distribution from a Uniform Process .....	35
Example 2.6	Exponential Distributions from Uniformly Distributed Process .....	37
Example 2.7	Uncorrelated Variables May not Be Independent .....	51
Example 3.8	An Epidemiology Example .....	61
Example 3.9	Particle Identification and Purity of a Sample .....	63
Example 3.10	Extreme Cases of Prior Beliefs .....	66
Example 3.11	Posterior for a Poisson Rate .....	71
Example 3.12	Posterior for Exponential Distribution .....	76
Example 4.13	Transition From Regular to ‘Unpredictable’ Sequences ....	82
Example 4.14	Extraction of an Exponential Random Variable .....	87
Example 4.15	Extraction of a Uniform Point on a Sphere .....	87
Example 4.16	Combining Different Monte Carlo Techniques .....	92
Example 5.17	A Very Simple Estimator in a Gaussian Case .....	101
Example 5.18	Estimators with Variance Below the Cramér–Rao Bound Are not Consistent .....	103
Example 5.19	Maximum Likelihood Estimate for an Exponential Distribution .....	112
Example 5.20	Bias of the Maximum Likelihood Estimate of a Gaussian Variance .....	113

Example 6.21	Reusing Multiple Times the Same Measurement Does not Improve a Combination.....	135
Example 7.22	Neyman Belt: Gaussian Case .....	146
Example 7.23	Application of the Clopper–Pearson Method.....	148
Example 9.24	Comparison of Multivariate Discriminators .....	199
Example 10.25	<i>p</i> -Value for a Poissonian Counting .....	206
Example 10.26	Can Frequentist and Bayesian Upper Limits Be ‘Unified’? .....	221
Example 10.27	Bump Hunting with the $L_{s+b}/L_b$ Test Statistic .....	232
Example 10.28	Adding Systematic Uncertainty with $L_{s+b}/L_b$ Approach...	236
Example 10.29	Bump Hunting with Profile Likelihood.....	239
Example 10.30	Simplified Look Elsewhere Calculation .....	245