

Data Analysis Techniques for Physical Scientists

Data Analysis Techniques for Physical Scientists is a comprehensive guide to data analysis techniques for physical scientists, providing a valuable resource for advanced undergraduate and graduate students, as well as seasoned researchers. The book begins with an extensive discussion of the foundational concepts and methods of probability and statistics under both the frequentist and Bayesian interpretations of probability. It next presents basic concepts and techniques used for measurements of particle production cross sections, correlation functions, and particle identification. Much attention is devoted to notions of statistical and systematic errors, beginning with intuitive discussions and progressively introducing the more formal concepts of confidence intervals, credible range, and hypothesis testing. The book also includes an in-depth discussion of the methods used to unfold or correct data for instrumental effects associated with measurement and process noise as well as particle and event losses, before ending with a presentation of elementary Monte Carlo techniques.

Claude A. Pruneau is a Professor of Physics at Wayne State University, from where he received the 2006 Excellence in Teaching Presidential award. He is also a member of the ALICE collaboration, and conducts an active research program in the study of the Quark Gluon Plasma produced in relativistic heavy ion collisions at the CERN Large Hadron Collider. He has worked as a Research Fellow at both Atomic Energy for Canada Limited and McGill University, and is a member of the American Physical Society, Canadian Association of Physicists and the Union of Concerned Scientists.

Data Analysis Techniques for Physical Scientists

CLAUDE A. PRUNEAU

Wayne State University, Michigan



CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom
One Liberty Plaza, 20th Floor, New York, NY 10006, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
4843/24, 2nd Floor, Ansari Road, Daryaganj, Delhi - 110002, India
79 Anson Road, #06-04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org

Information on this title: www.cambridge.org/9781108416788

DOI: 10.1017/9781108241922

© Cambridge University Press 2017

This publication is in copyright. Subject to statutory exception and to the provisions of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press.

First published 2017

Printed in the United Kingdom by TJ International Ltd. Padstow Cornwall

A catalogue record for this publication is available from the British Library

ISBN 978-1-108-41678-8 Hardback

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication, and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

To my son Blake

Contents

<i>Preface</i>	<i>page xi</i>
<i>How to Read This Book</i>	<i>xiii</i>
1 The Scientific Method	1
1.1 What the Brain Does	1
1.2 Critics of the Scientific Method	2
1.3 Falsifiability and Predictive Power	4
1.4 A Flawed but Effective Process!	5
1.5 Science as an Empirical Study of Nature	5
1.6 Defining Scientific Models	7
1.7 Challenges of the Scientific Method: Paradigm Shifts and Occam's Razor	10
1.8 To Err Is Human, to Control Errors, Science!	12
1.9 Goals, Structure, and Layout of This Book	13
Part I Foundation in Probability and Statistics	
2 Probability	17
2.1 Modeling Measurements: A Need for Probability	17
2.2 Foundation and Definitions	19
2.3 Frequentist and Bayesian Interpretations of Probabilities	29
2.4 Bayes' Theorem and Inference	33
2.5 Definition of Probability Distribution and Probability Density	35
2.6 Functions of Random Variables	39
2.7 PDF Characterization	41
2.8 Multivariate Systems	49
2.9 Moments of Multivariate PDFs	60
2.10 Characteristic and Moment-Generating Functions	66
2.11 Measurement Errors	73
2.12 Random Walk Processes	78
2.13 Cumulants	81
Exercises	84
3 Probability Models	88
3.1 Bernoulli and Binomial Distributions	88
3.2 Multinomial Distribution	96
3.3 Poisson Distribution	99

3.4	Uniform Distribution	106
3.5	Exponential Distribution	110
3.6	Gamma Distribution	113
3.7	Beta Distribution	115
3.8	Dirichlet Distributions	117
3.9	Gaussian Distribution	118
3.10	Multidimensional Gaussian Distribution	122
3.11	Log-Normal Distribution	124
3.12	Student's t -Distribution	126
3.13	Chi-Square Distribution	128
3.14	F -Distribution	131
3.15	Breit–Wigner (Cauchy) Distribution	132
3.16	Maxwell–Boltzmann Distribution	134
	Exercises	136
4	Classical Inference I: Estimators	139
4.1	Goals of Frequentist Inference	140
4.2	Population, Sample, and Sampling	140
4.3	Statistics and Estimators	144
4.4	Properties of Estimators	147
4.5	Basic Estimators	152
4.6	Histograms	159
4.7	Fisher Information	167
	Exercises	176
5	Classical Inference II: Optimization	178
5.1	The Method of Maximum Likelihood	179
5.2	The Method of Least-Squares	190
5.3	Determination of the Goodness-of-Fit	205
5.4	Extrapolation	206
5.5	Weighted Averages	207
5.6	Kalman Filtering	209
	Exercises	225
6	Classical Inference III: Confidence Intervals and Statistical Tests	227
6.1	Error Interval Estimation	227
6.2	Confidence Intervals with the Unified Approach	245
6.3	Confidence Intervals from ML and LS Fits	251
6.4	Hypothesis Testing, Errors, Significance Level, and Power	253
6.5	Test Properties	257
6.6	Commonly Used Tests	263
6.7	Test Optimization and the Neyman–Pearson Test	276
6.8	Appendix 1: Derivation of Student's t -Distribution	280
	Exercises	281

7 Bayesian Inference	284
7.1 Introduction	284
7.2 Basic Structure of the Bayesian Inference Process	286
7.3 Choosing a Prior	298
7.4 Bayesian Inference with Gaussian Noise	321
7.5 Bayesian Inference with Nonlinear Models and Non-Gaussian Processes	341
7.6 Optimization Techniques for Nonlinear Models	347
7.7 Model Comparison and Entity Classification	365
Exercises	385
Part II Measurement Techniques	
8 Basic Measurements	389
8.1 Basic Concepts and Notations	389
8.2 Particle Decays and Cross Sections	402
8.3 Measurements of Elementary Observables	409
8.4 Particle Identification	425
8.5 Detection of Short-Lived Particles	440
8.6 Searches, Discovery, and Production Limits	455
Exercises	458
9 Event Reconstruction	460
9.1 Event Reconstruction Overview	460
9.2 Track Reconstruction Techniques	471
9.3 Primary Vertex Reconstruction Techniques	490
9.4 Appendix 1: DCA Calculations	498
9.5 Appendix 2: Multiple Coulomb Scattering	501
Exercises	501
10 Correlation Functions	502
10.1 Extension of the Notion of Covariance	503
10.2 Correlation Function Cumulants	505
10.3 Semi-inclusive Correlation Functions	513
10.4 Factorial and Cumulant Moment-Generating Functions	515
10.5 Multivariate Factorial Moments	518
10.6 Correlation Functions of Nonidentical Particles	518
10.7 Charge-Dependent and Charge-Independent Correlation Functions	521
10.8 Generalized (Weighted) Correlation Functions	523
10.9 Autocorrelations and Time-Based Correlation Functions	524
Exercises	525
11 The Multiple Facets of Correlation Functions	526
11.1 Two-Particle Correlation Functions	526
11.2 Three-Particle Differential Correlation Functions	544

11.3	Integral Correlators	547
11.4	Flow Measurements	558
11.5	Appendix 1: Numerical Techniques Used in the Study of Correlation Functions	573
	Exercises	575
12	Data Correction Methods	577
12.1	Experimental Errors	577
12.2	Experimental Considerations	582
12.3	Signal Correction and Unfolding	587
12.4	Correcting Measurements of Correlation Functions	611
12.5	Systematic Errors	633
	Exercises	640
Part III Simulation Techniques		
13	Monte Carlo Methods	643
13.1	Basic Principle of Monte Carlo Methods	643
13.2	Monte Carlo Integration	644
13.3	Pseudorandom Number Generation	646
13.4	Selected Examples	657
	Exercises	661
14	Collision and Detector Modeling	663
14.1	Event Generators	663
14.2	Detector Simulation	678
	Exercises	686
	<i>References</i>	687
	<i>Index</i>	698

Preface

Physics students typically take a wide range of advanced classes in mechanics, electromagnetism, quantum mechanics, thermodynamics, and statistical mechanics, but sadly, receive only limited formal training in data analysis techniques. Most students in experimental physics indeed end up gleaning the required material by reading parts of a plurality of books and scientific articles. They typically end up knowing a lot about one particular analysis technique but relatively little about others. Paradoxically, modern experiments in particle and nuclear physics enable an amazingly wide range of very sophisticated measurements based on diverse analytical techniques. The end result is that beginning students may have a rather limited understanding of the many papers they become coauthors of by virtue of being members of a large scientific collaboration. After twenty years of teaching “physics” and carrying out research in heavy-ion physics, I figured I should make an effort to remedy this situation by creating a book that covers all the basic tools required in the data analysis of experiments at RHIC, the LHC, and other large experimental facilities.

This was a fairly ambitious project given that the range of techniques employed in today’s experiments is actually quite large and rather sophisticated. In the interest of full disclosure, I should state that the scope of the project changed several times, at times growing and at others shrinking. Eventually, I decided for a book in three parts covering (I) foundational concepts in probability and statistics, (II) basic and commonly used advanced measurement techniques, and (III) introductory techniques in Monte Carlo simulations targeted, mostly, toward the analysis and interpretation of experimental data. As such, it became impossible to present detailed descriptions of detector technologies or the physical principles they are based on. But as it turns out, high-quality data analyses are possible even if one is not familiar with the many technical details involved in the design or construction of detectors. Detector attributes relevant for data analyses can in general be reduced to a statement of a few essential properties, and it is thus possible to carry out quality analyses without a full knowledge of all aspects of a detector’s design and operation. I have thus opted to leave out detailed descriptions of detector technologies as well as particle interactions with matter and focus the discussion on some representative and illustrative examples of data calibration and analyses. Detailed discussions of detector technologies used in high-energy nuclear and particle physics may, however, be found in a plurality of graduate textbooks and technical texts. Additionally, I have also omitted few big and important topics such as interferometry (HBT), jet reconstruction, and neutral networks, for which very nice and comprehensive books or scientific reviews already exist.

Overall, this book essentially covers all basic techniques necessary for sound analyses and interpretation of experimental data. And, although it cannot cover all analysis techniques used by modern physicists, it lays a solid foundation in probability and statistics,

simulation techniques, and basic measurement methods, which should equip conscientious and dedicated students with the skill set they require for a successful career in experimental nuclear or particle physics, and such that they can explore more advanced techniques on their own.

I should note, in closing, that although this book targets primarily students in nuclear and particle physicists, it should, I believe, prove to be a useful introduction to data analysis for students working in other fields, including astronomy and basically all other areas of the physical sciences. It should also, I hope, provide a useful reference for more advanced and seasoned scientists.

I would like to express my sincere acknowledgments to the many people who, through discussions and advices, have helped shape this book. These include Monika Sharma, Rosie Reed, Robert Harr, Paul Karchin, and Sergei Voloshin, who through questions and comments have helped me plan or contributed various improvements to the book. I also wish to acknowledge the important contributions of several undergraduate and graduate students, most particularly Nick Elsey, Derek Everett, Derek Hazard, Ed Kramkowski, Jinjin Pan, Jon Troyer, and Chris Zin, who served as guinea-pigs for some fractions of the material. I am grateful to my colleagues Giovanni Bonvicini, from the CLEO Collaboration, for providing a Dalitz plot; Yuri Fisyak and Zhangbu Xu, from the STAR (Solenoidal Detector at Relativistic Heavy Ion Collider [RHIC]) Collaboration, for their contribution of a dE/dx plot; and my former postdoctoral student, Sidharth Prasad, for producing exemplars of unfolding. I also acknowledge use of several sets of results from the STAR collaboration, publicly available from the collaboration's website, for the generation of figures presenting examples of flow measurements and correlation functions. I am particularly indebted to colleagues Drs. Jean Barrette, Ron Belmont, Jana Bielcikova, Panos Christakoglou, Kolja Kauder, William Llope, Prabhat Pujahari, Sidharth Prasad, Joern Putschke, and William Zajc for their detailed reading and feedback on various sections of the book corresponding to their respective areas of expertise and interest. I also wish to acknowledge Ms. Heidi Kenaga and Ms. Theresa Kornak for their meticulous proofreading of the manuscript and for being so nice in correcting my Frenghish.

Finally, I wish to acknowledge that a large fraction of the graphs and figures featured in this book were created with ROOT, Keynote, and Graphic Converter. Several of the ROOT macros I wrote for the generation of figures will be made available at the book website.

Claude A. Pruneau

How to Read This Book

Not all students, instructors, and practitioners of the field of experimental physics may have the inclination, the time, or the need to study this book in its entirety. Indeed, only a selected few may have the opportunity to read the book from cover to cover. This should not be a problem, however, because the material is organized in large blocks that are reasonably self-sufficient, and ample references to earlier or upcoming chapters, as the case may be, are included in the narrative. The book also includes a number of specialized or in-depth topics that may be skipped in a first reading. Such topics include, for instance, the formal definition of probability in §2.2, the notion of Fisher information discussed in §4.7, the technique of Kalman filtering introduced in §5.6 and for which a detailed example of application is presented in §9.2.3, as well as discussions of track and vertex reconstruction presented in §§ 9.2 and 9.3. This said, the book is designed to progressively develop and approach topics, and it should then be possible to study the material in a variety of ways, adapting the depth and breadth of coverage. The following are recommended lists of chapters and sections that should be covered given specific and targeted needs.

- Introductory course in probability and statistics:
Chapters 2 (§§2.1, 2.3–2.11), 3 (§§3.1–3.13), 4 (§§4.1–4.6), 5 (§§5.1–5.5), 6 (§§6.1–6.6), 7 (§§7.1, 7.2, 7.4, 7.7), 13
- Advanced course in probability and statistics:
Chapters 1, 2, 3, 4, 5, 6, 7, 13
- Introductory course in data analysis techniques (one semester):
Chapters 1, 2 (§§2.1, 2.3–2.11), 3 (§§3.1–3.13), 4 (§§4.1–4.6), 5 (§§5.1–5.5), 6 (§§6.1–6.6), 8 (§§8.1–8.6), 9 (§§9.1, 9.2), 13
- Advanced course in data analysis techniques (two semesters):
Chapters 1, 2 (§§2.1, 2.3–2.11), 3 (§§3.1–3.13), 4 (§§4.1–4.6), 5 (§§5.1–5.6), 6 (§§6.1–6.7), 7, 8 (§§8.1–8.6), 9 (§§9.1, 9.2), 12, 13, 14
- Course on correlation functions (one semester):
Chapters 2 (§§2.5–2.13), 4 (§§4.3, 4.5, 4.6), 10, 11, 12, 13

Of course, instructors using this book should feel free to select and change the order of topics to suit their specific needs. For instance, Monte Carlo methods are formally introduced in Chapter 13 but it is often useful and inconvenient to use and discuss some of these concepts along with materials of the early chapters (e.g., Chapters 2–7).