Modern Statistical Methods for Astronomy

With **R** Applications

ERIC D. FEIGELSON
Pennsylvania State University

G. JOGESH BABU
Pennsylvania State University

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town,
Singapore, São Paulo, Delhi, Mexico City

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521767279

© E. D. Feigelson and G. J. Babu 2012

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 2012

Printed in the United Kingdom at the University Press, Cambridge

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

Library of Congress Catalog in Publication data
Feigelson, Eric D.
Modern statistical methods for astronomy : with R applications /
Eric D. Feigelson, G. Jogesh Babu.
p. cm.
ISBN 978-0-521-76727-9 (hardback
1. Statistical astronomy. I. Babu, Gutti Jogesh, 1949– II. Title.
QB149.F45 2012
520.72′7 – dc23 2012009113

ISBN 978-0-521-76727-9 Hardback

Additional resources for this publication: www.cambridge.org/msma

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.

Contents

Preface *page* xv

1 Introduction 1

1.1 The role of statistics in astronomy 1

1.1.1 Astronomy and astrophysics 1

1.1.2 Probability and statistics 3

1.1.3 Statistics and science 4

1.2 History of statistics in astronomy 6

1.2.1 Antiquity through the Renaissance 6

1.2.2 Foundations of statistics in celestial mechanics 7

1.2.3 Statistics in twentieth-century astronomy 8

1.3 Recommended reading 10

2 Probability 13

2.1 Uncertainty in observational science 13

2.2 Outcome spaces and events 14

2.3 Axioms of probability 15

2.4 Conditional probabilities 17

2.4.1 Bayes' theorem 18

2.4.2 Independent events 19

2.5 Random variables 20

2.5.1 Density and distribution functions 21

2.5.2 Independent and identically distributed r.v.s. 24

2.6 Quantile function 25

2.7 Discrete distributions 26

2.8 Continuous distributions 27

2.9 Distributions that are neither discrete nor continuous 29

2.10 Limit theorems 30

2.11 Recommended reading 30

2.12 **R** applications 31

3 Statistical inference 35

3.1 The astronomical context 35

3.2 Concepts of statistical inference 36

3.3 Principles of point estimation 38

3.4 Techniques of point estimation 40

3.4.1 Method of moments 41

3.4.2 Method of least squares 42

3.4.3 Maximum likelihood method 43

3.4.4 Confidence intervals 45

3.4.5 Calculating MLEs with the EM algorithm 47

3.5 Hypothesis testing techniques 48

3.6 Resampling methods 52

3.6.1 Jackknife 52

3.6.2 Bootstrap 54

3.7 Model selection and goodness-of-fit 57

3.7.1 Nonparametric methods for goodness-of-fit 58

3.7.2 Likelihood-based methods for model selection 60

3.7.3 Information criteria for model selection 61

3.7.4 Comparing different model families 62

3.8 Bayesian statistical inference 63

3.8.1 Inference for the binomial proportion 64

3.8.2 Prior distributions 65

3.8.3 Inference for Gaussian distributions 67

3.8.4 Hypotheses testing and the Bayes factor 69

3.8.5 Model selection and averaging 70

3.8.6 Bayesian computation 71

3.9 Remarks 72

3.10 Recommended reading 73

3.11 **R** applications 74

4 Probability distribution functions 76

4.1 Binomial and multinomial 76

4.1.1 Ratio of binomial random variables 79

4.2 Poisson 80

4.2.1 Astronomical context 80

4.2.2 Mathematical properties 81

4.2.3 Poisson processes 83

4.3 Normal and lognormal 85

4.4 Pareto (power-law) 87

4.4.1 Least-squares estimation 89

4.4.2 Maximum likelihood estimation 90

4.4.3 Extensions of the power-law 91

4.4.4 Multivariate Pareto 92

4.4.5 Origins of power-laws 93

4.5 Gamma 94

4.6 Recommended reading 96

4.7 **R** applications 96

4.7.1 Comparing Pareto distribution estimators 97

4.7.2 Fitting distributions to data 101
4.7.3 Scope of distributions in **R** and **CRAN** 103

5 Nonparametric statistics 105
5.1 The astronomical context 105
5.2 Concepts of nonparametric inference 106
5.3 Univariate problems 107
5.3.1 Kolmogorov–Smirnov and other e.d.f. tests 107
5.3.2 Robust statistics of location 110
5.3.3 Robust statistics of spread 111
5.4 Hypothesis testing 111
5.4.1 Sign test 112
5.4.2 Two-sample and *k*-sample tests 112
5.5 Contingency tables 113
5.6 Bivariate and multivariate tests 115
5.7 Remarks 116
5.8 Recommended reading 117
5.9 **R** applications 117
5.9.1 Exploratory plots and summary statistics 117
5.9.2 Empirical distribution and quantile functions 121
5.9.3 Two-sample tests 124
5.9.4 Contingency tables 125
5.9.5 Scope of nonparametrics in **R** and **CRAN** 127

6 Data smoothing: density estimation 128
6.1 The astronomical context 128
6.2 Concepts of density estimation 128
6.3 Histograms 129
6.4 Kernel density estimators 131
6.4.1 Basic properties 131
6.4.2 Choosing bandwidths by cross-validation 132
6.4.3 Multivariate kernel density estimation 133
6.4.4 Smoothing with measurement errors 134
6.5 Adaptive smoothing 134
6.5.1 Adaptive kernel estimators 134
6.5.2 Nearest-neighbor estimators 135
6.6 Nonparametric regression 136
6.6.1 Nadaraya–Watson estimator 136
6.6.2 Local regression 137
6.7 Remarks 138
6.8 Recommended reading 138
6.9 **R** applications 139
6.9.1 Histogram, quantile function and measurement errors 139
6.9.2 Kernel smoothers 140

6.9.3 Nonparametric regressions 144
6.9.4 Scope of smoothing in **R** and **CRAN** 148

7 Regression 150
7.1 Astronomical context 150
7.2 Concepts of regression 151
7.3 Least-squares linear regression 154
7.3.1 Ordinary least squares 154
7.3.2 Symmetric least-squares regression 155
7.3.3 Bootstrap error analysis 156
7.3.4 Robust regression 158
7.3.5 Quantile regression 160
7.3.6 Maximum likelihood estimation 161
7.4 Weighted least squares 162
7.5 Measurement error models 164
7.5.1 Least-squares estimators 166
7.5.2 SIMEX algorithm 168
7.5.3 Likelihood-based estimators 169
7.6 Nonlinear models 169
7.6.1 Poisson regression 170
7.6.2 Logistic regression 171
7.7 Model validation, selection and misspecification 172
7.7.1 Residual analysis 173
7.7.2 Cross-validation and the bootstrap 175
7.8 Remarks 176
7.9 Recommended reading 177
7.10 **R** applications 177
7.10.1 Linear modeling 179
7.10.2 Generalized linear modeling 181
7.10.3 Robust regression 182
7.10.4 Quantile regression 183
7.10.5 Nonlinear regression of galaxy surface brightness profiles 184
7.10.6 Scope of regression in **R** and **CRAN** 189

8 Multivariate analysis 190
8.1 The astronomical context 190
8.2 Concepts of multivariate analysis 191
8.2.1 Multivariate distances 192
8.2.2 Multivariate normal distribution 194
8.3 Hypothesis tests 195
8.4 Relationships among the variables 197
8.4.1 Multiple linear regression 197
8.4.2 Principal components analysis 199
8.4.3 Factor and canonical correlation analysis 200

8.4.4 Outliers and robust methods 201
8.4.5 Nonlinear methods 202
8.5 Multivariate visualization 203
8.6 Remarks 204
8.7 Recommended reading 205
8.8 **R** applications 206
8.8.1 Univariate tests of normality 206
8.8.2 Preparing the dataset 208
8.8.3 Bivariate relationships 209
8.8.4 Principal components analysis 212
8.8.5 Multiple regression and MARS 214
8.8.6 Multivariate visualization 216
8.8.7 Interactive graphical displays 217
8.8.8 Scope of multivariate analysis **R** and **CRAN** 220

9 Clustering, classification and data mining 222
9.1 The astronomical context 222
9.2 Concepts of clustering and classification 224
9.2.1 Definitions and scopes 224
9.2.2 Metrics, group centers and misclassifications 225
9.3 Clustering 226
9.3.1 Agglomerative hierarchical clustering 226
9.3.2 *k*-means and related nonhierarchical partitioning 228
9.4 Clusters with substructure or noise 229
9.5 Mixture models 231
9.6 Supervised classification 232
9.6.1 Multivariate normal clusters 232
9.6.2 Linear discriminant analysis and its generalizations 233
9.6.3 Classification trees 234
9.6.4 Nearest-neighbor classifiers 236
9.6.5 Automated neural networks 237
9.6.6 Classifier validation, improvement and fusion 238
9.7 Remarks 239
9.8 Recommended reading 241
9.9 **R** applications 242
9.9.1 Unsupervised clustering of COMBO-17 galaxies 242
9.9.2 Mixture models 246
9.9.3 Supervised classification of SDSS point sources 250
9.9.4 LDA, *k*-nn and ANN classification 251
9.9.5 CART and SVM classification 255
9.9.6 Scope of **R** and **CRAN** 259

10 Nondetections: censored and truncated data 261
10.1 The astronomical context 261

10.2 Concepts of survival analysis 263
10.3 Univariate datasets with censoring 266
10.3.1 Parametric estimation 266
10.3.2 Kaplan–Meier nonparametric estimator 268
10.3.3 Two-sample tests 269
10.4 Multivariate datasets with censoring 271
10.4.1 Correlation coefficients 271
10.4.2 Regression models 272
10.5 Truncation 274
10.5.1 Parametric estimation 275
10.5.2 Nonparametric Lynden-Bell–Woodroofe estimator 275
10.6 Remarks 277
10.7 Recommended reading 278
10.8 **R** applications 279
10.8.1 Kaplan–Meier estimator 279
10.8.2 Two-sample tests with censoring 281
10.8.3 Bivariate and multivariate problems with censoring 284
10.8.4 Lynden-Bell–Woodroofe estimator for truncation 287
10.8.5 Scope of censoring and truncation in **R** and **CRAN** 290

11 Time series analysis 292
11.1 The astronomical context 292
11.2 Concepts of time series analysis 294
11.3 Time-domain analysis of evenly spaced data 296
11.3.1 Smoothing 296
11.3.2 Autocorrelation and cross-correlation 297
11.3.3 Stochastic autoregressive models 298
11.3.4 Regression for deterministic models 301
11.4 Time-domain analysis of unevenly spaced data 302
11.4.1 Discrete correlation function 302
11.4.2 Structure function 304
11.5 Spectral analysis of evenly spaced data 304
11.5.1 Fourier power spectrum 305
11.5.2 Improving the periodogram 307
11.6 Spectral analysis of unevenly spaced data 308
11.6.1 Lomb–Scargle periodogram 308
11.6.2 Non-Fourier periodograms 310
11.6.3 Statistical significance of periodogram peaks 312
11.6.4 Spectral analysis of event data 313
11.6.5 Computational issues 314
11.7 State-space modeling and the Kalman filter 315
11.8 Nonstationary time series 317
11.9 $1/f$ noise or long-memory processes 319
11.10 Multivariate time series 322

11.11 Remarks 323
11.12 Recommended reading 324
11.13 **R** applications 325
11.13.1 Exploratory time series analysis 326
11.13.2 Spectral analysis 329
11.13.3 Modeling as an autoregressive process 330
11.13.4 Modeling as a long-memory process 333
11.13.5 Wavelet analysis 334
11.13.6 Scope of time series analysis in **R** and **CRAN** 336

12 Spatial point processes 337
12.1 The astronomical context 337
12.2 Concepts of spatial point processes 338
12.3 Tests of uniformity 340
12.4 Spatial autocorrelation 341
12.4.1 Global measures of spatial autocorrelation 341
12.4.2 Local measures of spatial autocorrelation 343
12.5 Spatial interpolation 344
12.6 Global functions of clustering 346
12.6.1 Cumulative second-moment measures 346
12.6.2 Two-point correlation function 348
12.7 Model-based spatial analysis 351
12.7.1 Models for galaxy clustering 351
12.7.2 Models in geostatistics 353
12.8 Graphical networks and tessellations 354
12.9 Points on a circle or sphere 355
12.10 Remarks 357
12.11 Recommended reading 358
12.12 **R** applications 359
12.12.1 Characterization of autocorrelation 361
12.12.2 Variogram analysis 362
12.12.3 Characterization of clustering 364
12.12.4 Tessellations 368
12.12.5 Spatial interpolation 370
12.12.6 Spatial regression and modeling 373
12.12.7 Circular and spherical statistics 374
12.12.8 Scope of spatial analysis in **R** and **CRAN** 377

Appendix A Notation and acronyms 379

Appendix B Getting started with R 382
B.1 History and scope of **R/CRAN** 382
B.2 Session environment 382
B.3 **R** object classes 385

B.4 Basic operations on classes 386
B.5 Input/output 388
B.6 A sample **R** session 389
B.7 Interfaces to other programs and languages 394
B.8 Computational efficiency 394
B.9 Learning more about **R** 397
B.10 Recommended reading 398

Appendix C Astronomical datasets 399
C.1 Asteroids 400
C.2 Protostar populations 402
C.3 Globular cluster magnitudes 403
C.4 Stellar abundances 405
C.5 Galaxy clustering 406
C.6 Hipparcos stars 408
C.7 Globular cluster properties 410
C.8 SDSS quasars 411
C.9 SDSS point sources 413
C.10 Galaxy photometry 419
C.11 Elliptical galaxy profiles 420
C.12 X-ray source variability 421
C.13 Sunspot numbers 422
C.14 Exoplanet orbits 423
C.15 Kepler stellar light curves 425
C.16 Sloan Digital Sky Survey 428
C.17 Fermi gamma-ray light curves 430
C.18 Swift gamma-ray bursts 432

References 434
Subject index 462
R and CRAN commands 470

The color plates are to be found between pages 398 and 399.

Preface

Motivation and goals

For many years, astronomers have struggled with the application of sophisticated statistical methodologies to analyze their rich datasets and address complex astrophysical problems. On one hand, at least in the United States, astronomers receive little or no formal training in statistics. The traditional method of education has been informal exposure to a few familiar methods during early research experiences. On the other hand, astronomers correctly perceive that a vast world of applied mathematical and statistical methodologies has emerged in recent decades. But systematic, broad training in modern statistical methods has not been available to most astronomers.

This volume seeks to address this problem at three levels. First, we present fundamental principles and results of broad fields of statistics applicable to astronomical research. The material is roughly at a level of advanced undergraduate courses in statistics. We also outline some recent advanced techniques that may be useful for astronomical research to give a flavor of the breadth of modern methodology. It is important to recognize that we give only incomplete introductions to the fields, and we guide the astronomer towards more complete and authoritative treatments.

Second, we present tutorials on the application of both simple and more advanced methods applied to contemporary astronomical research datasets using the **R** statistical software package. **R** has emerged in recent years as the most versatile public-domain statistical software environment for researchers in many fields. In addition to a coherent language for data analysis and common statistical tools, over 3000 packages have been added for advanced analyses in the **CRAN** archive. We have culled these packages for functionalities that may be useful to astronomers. **R** can also be linked to other analysis systems and languages such as C, FORTRAN and Python, so that legacy codes can be included in an **R**-based analysis and *vice versa*.

Third, we hope the book communicates to astronomers our enthusiasm for statistics as a substantial and fascinating intellectual enterprise. Just as astronomers use the latest engineering to build their telescopes and apply advanced physics to interpret cosmic phenomena, they can benefit from exploring the many roads of analyzing and interpreting data through modern statistical analysis.

Another important purpose of this volume is to give astronomers and other physical scientists a bridge to the vast library of specialized texts and monographs in statistics and allied fields. We strongly encourage researchers who are engaged in statistical data analysis to read more detailed treatments in the ‘Recommended reading’ at the end of each chapter; they are carefully chosen from many available volumes. Most of the material in the book which is not specifically referenced in the text is presented in more detail in these recommended readings. To further this goal, the present book does not shy away from technical language that, though unfamiliar in the astronomical community, is critical for further learning from the statistical literature. For example, the astronomers' "upper limits" are "left-censored data points", a "power-law distribution" is a "Pareto distribution", and "$1/f$ noise" is a "long-memory process". The text make these connections between the languages of astronomy and statistics, and the comprehensive index can assist the reader in finding material in both languages.

The reader may find the appendices useful. An introduction to **R** is given in Appendix B. It includes an overview of the programming language and an outline of its statistical functionalities, including the many **CRAN** packages. **R** applications to astronomical datasets are given at the end of each chapter which implement methods discussed in the text. Appendix C presents 18 astronomical datasets illustrating the range of statistical challenges that arise in contemporary research. The full datasets and **R scripts** are available online at http://astrostatistics.psu.edu/MSMA. Readers can thus easily reproduce the **R** results in the book.

In this volume, we do not present mathematical proofs underlying statistical results, and we give only brief outlines of a few computational algorithms. We do not review research at the frontiers of astrostatistics, except for a few topics where astronomers have contributed critically important methodology (such as the treatment of truncated data and irregularly spaced time series). Only a small fraction of the many methodological studies in the recent astronomical literature are mentioned. Some fields of applied statistics useful for astronomy (such as wavelet analysis and image processing) are covered only briefly. Finally, only ∼2500 **CRAN** packages were examined for possible inclusion in the book; roughly one new package is added every day and many others are extended.

Audience

The main audience envisioned for this volume is graduate students and researchers in observational astronomy. We hope it serves both as a textbook in a course on data analysis or astrostatistics, and as a reference book to be consulted as specific research problems are encountered. Researchers in allied fields of physical science, such as high-energy physics and Earth sciences, may also find portions of the volume helpful. Statisticians can see how existing methods relate to questions in astronomy, providing background for astrostatistical research initiatives.

Our presentation assumes that the reader has a background in basic linear algebra and calculus. Familiarity of elementary statistical methods commonly used in the physical sciences is also useful; this preparatory material is covered in volumes such as Bevington & Robinson (2002) and Cowan (1998).

Outline and classroom use

The introduction (Chapter 1) reviews the long historical relationship between astronomy and statistics and philosophical discussions of the relationship between statistical and scientific inference. We then start with probability theory and proceed to lay foundations of statistical inference: hypothesis testing, estimation, modeling, resampling and Bayesian inference

(Chapters 2 and 3). Probability distributions are discussed in Chapter 4 and nonparametric statistics are covered in Chapter 5.

The volume proceeds to various fields of applied statistics that rest on these foundations. Data smoothing is covered in Chapters 5 and 6. Regression is discussed in Chapter 7, followed by analysis and classification of multivariate data (Chapters 8 and 9). Treatments of nondetections are covered in Chapter 10, followed by the analysis of time-variable astronomical phenomena in Chapter 11. Chapter 12 considers spatial point processes. The book ends with appendices introducing the **R** software environment and providing astronomical datasets illustrative of a variety of statistical problems.

We can make some recommendation regarding classroom use. The first part of a semester course in astrostatistics for astronomy students would be devoted to the principles of statistical inference in Chapters 1–4 and learning the basics of **R** in Appendix B. The second part of the semester would be topics of applied statistical methodology selected from Chapters 5–12. We do not provide predefined student exercises with definitive answers, but rather encourage both instructors and students to develop open-ended explorations of the contemporary astronomical datasets based on the **R** tutorials distributed throughout the volume. Suggestions for both simple and advanced problems are given in the dataset presentations (Appendix C).

Astronomical datasets and R scripts

The datasets and **R** scripts in the book can be downloaded from Penn State's Center for Astrostatistics at http://astrostatistics.psu.edu/MSMA. The **R** scripts are self-contained; simple cut-and-paste will ingest the datasets, perform the statistical operations, and produce tabular and graphical results.

Extensive resources to pursue issues discussed in the book are available on-line. The **R** system can be downloaded from http://www.r-project.org and **CRAN** packages are installed on-the-fly within an **R** session. The primary astronomy research literature, including full-text articles, is available through the NASA–Smithsonian *Astrophysics Data System* (http://adswww.harvard.edu). Thousands of astronomical datasets are available from the Vizier service at the Centre des Données Stellaires (http://vizier.u-strasbg.fr) and the emerging *International Virtual Observatory Alliance* (http://ivo.net). The primary statistical literature can be accessed through MathSciNet (http://www.ams.org/mathscinet/) provided by the American Mathematical Society. Considerable statistical information is available on Wikipedia (http://en.wikipedia.org/wiki/Index_of_statistics_articles). Astronomers should note, however, that the best way to learn statistics is often through textbooks and monographs written by statisticians, such as those in the recommended reading.

Acknowledgements

This book emerged from 25 years of discussion and collaboration between astronomers and statisticians at Penn State under the auspices of the Center for Astrostatistics. The volume particularly benefited from the lectures and tutorials developed for the *Summer Schools*

in Statistics for Astronomers since 2005 and taught at Penn State and Bangalore's Indian Institute of Astrophysics. We are grateful to our dozens of statistician colleagues who have taught at the *Summer Schools in Statistics for Astronomers* for generously sharing their knowledge and perspectives. David Hunter and Arnab Chakraborty developed **R** tutorials for astronomers. Donald Percival generously gave detailed comments on the time series analysis chapter. We are grateful to Nancy Butkovich and her colleagues for providing excellent library services. Finally, we acknowledge the National Science Foundation, National Aeronautics and Space Administration, and the Eberly College of Science for supporting astrostatistics at Penn State over many years.

Eric D. Feigelson
G. Jogesh Babu
Center for Astrostatistics
Pennsylvania State University
University Park, PA, U.S.A.