Exponential Families in Theory and Practice

During the past half-century, exponential families have attained a position at the center of parametric statistical inference. Theoretical advances have been matched, and more than matched, in the world of applications, where logistic regression by itself has become the go-to methodology in medical statistics, computer-based prediction algorithms, and the social sciences. This book is based on a one-semester graduate course for first year Ph.D. and advanced master's students. After presenting the basic structure of univariate and multivariate exponential families, their application to generalized linear models including logistic and Poisson regression is described in detail, emphasizing geometrical ideas, computational practice, and the analogy with ordinary linear regression. Connections are made with a variety of current statistical methodologies: missing data, survival analysis and proportional hazards, false discovery rates, bootstrapping, and empirical Bayes analysis. The book connects exponential family theory with its applications in a way that doesn't require advanced mathematical preparation.

BRADLEY EFRON is Professor Emeritus of Statistics and Biomedical Data Science at Stanford University. He is the inventor of the bootstrap method for assessing statistical accuracy. He has published extensively on statistical theory and its applications, with particular attention to exponential families. A MacArthur fellow, he is a member of the National Academy of Sciences. He received the National Medal of Science in 2007.

INSTITUTE OF MATHEMATICAL STATISTICS TEXTBOOKS

IMS Textbooks give introductory accounts of topics of current concern suitable for advanced courses at master's level, for doctoral students and for individual study. They are typically shorter than a fully developed textbook, often arising from material created for a topical course. Lengths of 100–290 pages are envisaged. The books typically contain exercises.

In collaboration with the International Society for Bayesian Analysis (ISBA), selected volumes in the IMS Textbooks series carry the "with ISBA" designation at the recommendation of the ISBA editorial representative.

Other Books in the Series (*with ISBA)

1. *Probability on Graphs*, by Geoffrey Grilmmett
2. *Stochastic Networks*, by Frank Kelly and Elena Yudovina
3. *Bayesian Filtering and Smoothing*, by Simo Särkkä
4. *The Surprising Mathematics of Longest Increasing Subsequences*, by Dan Romik
5. *Noise Sensitivity of Boolean Functions and Percolation*, by Christophe Garban and Jeffrey E. Steif
6. *Core Statistics*, by Simon N. Wood
7. *Lectures on the Poisson Process*, by Günter Last and Mathew Penrose
8. *Probability on Graphs (Second Edition)*, by Geoffrey Grimmett
9. *Introduction to Malliavin Calculus*, by David Nualart and Eulália Nualart
10. *Applied Stochastic Differential Equations*, by Simo Särkkä and Arno Solin
11. **Computational Bayesian Statistics*, by M. Antónia Amaral Turkman, Carlos Daniel Paulino, and Peter Müller
12. *Statistical Modelling by Exponential Families*, by Rolf Sundberg
13. *Two-Dimensional Random Walk: From Path Counting to Random Interlacements*, by Serguei Popov
14. *Scheduling and Control of Queueing Networks*, by Gideon Weiss
15. *Principles of Statistical Analysis: Learning from Randomized Experiments*, by Ery Arias-Castro
16. *Exponential Families in Theory and Practice*, by Bradley Efron

Exponential Families in Theory and Practice

BRADLEY EFRON
Stanford University

CAMBRIDGE UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment, a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781108488907

DOI: 10.1017/9781108773157

First published 2023

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

ISBN 978-1-108-48890-7 Hardback
ISBN 978-1-108-71566-9 Paperback

Contents

Preface

Exponential Families in Theory and Practice is based on my notes for a graduate course designed for first-year Ph.D. and advanced master's degree students in the Statistics Department at Stanford. The course and the book focus on the elegant structure of exponential families, and how exponential family methods have transformed statistical applications in the age of high-speed computing.

Parts 1, 2, and 3 concern the basic ideas of univariate and multivariate exponential families, and their use in generalized linear models, particularly logistic and Poisson regression, the mainstays of modern applications in a variety of fields. The three parts can be covered in about twenty 50-minute lectures, leaving ten lectures for selections from Parts 4 and 5 in a one-quarter course, or fifteen in a semester. Applied topics touch on several statistical success stories: survival analysis and proportional hazards, empirical Bayes, missing data, and false discovery rates.

Homework problems, integrated into the text rather than gathered at the end, play an important role in getting the material across. For the most part the problems aren't very difficult, with the majority chosen to augment points raised in the lecture. Their main role is to help students incorporate the ideas rather than just hear them. Each week I usually assigned four or five homework problems to be turned in, and allowed students to work together on them.

Computational exercises utilize the programming language R, which is also used occasionally in the text to convey specific algorithmic details. Data sets appearing in the text are available from the author's website.

About the mathematical level: I have tried to keep this as low as possible consonant with the subject's needs. Asymptotic arguments are mostly absent, and there are almost no proofs except those that are vital to understanding the statistical points being made. A good background in multivariable calculus, linear algebra, and probability is sufficient mathematical background for the book. Exponential family theory has a strong geometri-

cal aspect, and, whenever possible, I have substituted geometry for algebra and figures for equations.

The physics profession has an honored cohort of practitioners called "phenomenologists" who work to connect theory with applications. In that spirit, the title *Exponential Families in Theory and Practice* could be better amended to *Exponential Families* Between *Theory and Practice*. My goal was to link the powerful theory of exponential families with the modern world of statistical applications, and I hope the book will be successful in that role for both teachers and students.

Acknowledgments

The material in this book accrued over fifty years of teaching, during which time the Stanford Statistics graduate students were almost always good sports and keen critics. My associate Cindy Kirby did heroic work as editor, compositor, and occasional artist in turning messy notes into the volume you are holding. My thanks also to my Cambridge University Press editors Lauren Cowles and Diana Gilooly for their kind support during the long process of publication.

Introduction

Some great ideas are born in a flash of inspiration, perhaps announced to the world by a pathbreaking paper. R. A. Fisher's 1925 article on maximum likelihood estimation is a classic example. Nothing at all like that happened with exponential families. The theory accrued slowly over a period extending roughly between 1932 and 1970. Applications lagged behind, a turning point being the advent of logistic regression and McCullagh and Nelder's 1983 book on generalized linear models.

A salient fact is that no one person is credited with the development of exponential families, though it will be clear from these notes that Fisher's work was instrumental. The name "exponential familes" is comparatively recent. Until the late 1950s they were often referred to as "Koopman–Darmois–Pitman" families (crediting three prominent statisticians working separately in three different countries); the awkward nomenclature suggests only minor importance being attached to the ideas.

Figure 1 gives a rough schematic history of Twentieth Century statistics. The inner circle represents normal theory, the preferred venue of classical methodology. Exact inference – t tests, F tests, chi-squared statistics, ANOVA, multivariate analysis – was feasible inside the circle. Outside the circle was a general theory based on large-sample asymptotic approximations involving Taylor series, Edgeworth expansions, and the central limit theorem. A few special exact results lay outside the normal circle, relating to especially tractable distributions such as the binomial, Poisson, gamma and beta families. These are the figure's green stars. A happy surprise, though a slowly emerging one beginning in the 1930s, was that the special cases were all examples of a powerful general construction, *exponential families*, the intermediate circle in Figure 1. Within this circle, "almost exact" inferential calculations are possible, where any necessary approximations can be pictured in simple geometric diagrams. Such diagrams play a major role in what follows.

Two complementary types of mathematical development can be labeled