Springer Series in Statistics

Springer Series in Statistics

Alho/Spencer: Statistical Demography and Forecasting
Andersen/Borgan/Gill/Keiding: Statistical Models Based on Counting Processes
Atkinson/Riani: Robust Diagnostic Regression Analysis
Atkinson/Riani/Ceriloi: Exploring Multivariate Data with the Forward Search
Berger: Statistical Decision Theory and Bayesian Analysis, 2nd edition
Borg/Groenen: Modern Multidimensional Scaling: Theory and Applications, 2nd edition
Brockwell/Davis: Time Series: Theory and Methods, 2nd edition
Bucklew: Introduction to Rare Event Simulation
Cappé/Moulines/Rydén: Inference in Hidden Markov Models
Chan/Tong: Chaos: A Statistical Perspective
Chen/Shao/Ibrahim: Monte Carlo Methods in Bayesian Computation
Coles: An Introduction to Statistical Modeling of Extreme Values
Devroye/Lugosi: Combinatorial Methods in Density Estimation
Diggle/Ribeiro: Model-based Geostatistics
Dudoit/Van der Laan: Multiple Testing Procedures with Applications to Genomics
Efromovich: Nonparametric Curve Estimation: Methods, Theory, and Applications
Eggermont/LaRiccia: Maximum Penalized Likelihood Estimation, Volume I: Density Estimation
Fahrmeir/Tutz: Multivariate Statistical Modeling Based on Generalized Linear Models, 2nd edition
Fan/Yao: Nonlinear Time Series: Nonparametric and Parametric Methods
Ferraty/Vieu: Nonparametric Functional Data Analysis: Theory and Practice
Ferreira/Lee: Multiscale Modeling: A Bayesian Perspective
Fienberg/Hoaglin: Selected Papers of Frederick Mosteller
Frühwirth-Schnatter: Finite Mixture and Markov Switching Models
Ghosh/Ramamoorthi: Bayesian Nonparametrics
Glaz/Naus/Wallenstein: Scan Statistics
Good: Permutation Tests: Parametric and Bootstrap Tests of Hypotheses, 3rd edition
Gouriéroux: ARCH Models and Financial Applications
Gu: Smoothing Spline ANOVA Models
Gyöfi/Kohler/Krzyźak/Walk: A Distribution-Free Theory of Nonparametric Regression
Haberman: Advanced Statistics, Volume I: Description of Populations
Hall: The Bootstrap and Edgeworth Expansion
Härdle: Smoothing Techniques: With Implementation in S
Harrell: Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis
Hart: Nonparametric Smoothing and Lack-of-Fit Tests
Hastie/Tibshirani/Friedman: The Elements of Statistical Learning: Data Mining, Inference, and Prediction
Hedayat/Sloane/Stufken: Orthogonal Arrays: Theory and Applications
Heyde: Quasi-Likelihood and Its Application: A General Approach to Optimal Parameter Estimation
Huet/Bouvier/Poursat/Jolivet: Statistical Tools for Nonlinear Regression: A Practical Guide with S-PLUS and R Examples, 2nd edition
Iacus: Simulation and Inference for Stochastic Differential Equations

(continued after index)

Friedrich Liese · Klaus-J. Miescke

Statistical Decision Theory

Estimation, Testing, and Selection

Springer

Friedrich Liese
Universität Rostock
Institut für Mathematik
Universitätsplatz 1
18051 Rostock
Germany
friedrich.liese@mathematik.uni-rostock.de

Klaus-J. Miescke
Department of Mathematics, Statistics
& Computer Science
University of Illinois at Chicago
851 South Morgan Street
Chicago IL 60607-7045
USA
klaus@uic.edu

ISBN: 978-0-387-73193-3 e-ISBN: 978-0-387-73194-0
DOI: 10.1007/978-0-387-73194-0

Library of Congress Control Number: 2008924221

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

To Gabi and Madelyn

Preface

This monograph is written for advanced Master's students, Ph.D. students, and researchers in mathematical statistics and decision theory. It should be useful not only as a basis for graduate courses, seminars, Ph.D. programs, and self-studies, but also as a reference tool.

At the very least, readers should be familiar with basic concepts covered in both advanced undergraduate courses on probability and statistics and introductory graduate-level courses on probability theory, mathematical statistics, and analysis. Most statements and proofs appear in a form where standard arguments from measure theory and analysis are sufficient. When additional information is necessary, technical tools, additional measure-theoretic facts, and advanced probabilistic results are presented in condensed form in an appendix. In particular, topics from measure theory and from the theory of weak convergence of distributions are treated in detail with reference to modern books on probability theory, such as Billingsley (1968), Kallenberg (1997, 2002), and Dudley (2002).

Building on foundational knowledge, this book acquaints readers with the concepts of classical finite sample size decision theory and modern asymptotic decision theory in the sense of LeCam. To this end, systematic applications to the fields of parameter estimation, testing hypotheses, and selection of populations are included. Some of the problems contain additional information in order to round off the results, whereas other problems, equipped with solutions, have a more technical character. The latter play the role of auxiliary results and as such they allow readers to become familiar with the advanced techniques of mathematical statistics.

The central theme of this book is what optimal decisions are in general and in specific decision problems, and how to derive them. Optimality is understood in terms of the expected loss, i.e. the risk, or some functional of it. In this regard estimators, tests, and selection rules are initially considered in the book side by side, and then individually in the last three chapters.

Originally we were motivated to write this book by the lack of any noticeable coverage of selection rules in books on decision theory. In over more

than 50 years' worth of scholarship, the majority of the over 1000 published articles on selection rules do not utilize a rigorous decision-theoretic approach. Instead, many articles on selection rules restrict themselves to a specific parametric family, propose an ad hoc rule, study its performance characteristics, and (at the very best) compare its performance with another competing selection rule. By contrast, this book offers a fuller point of view, and the last chapter provides a thorough presentation of optimal selection rule theory.

Two other justifications for including selection theory are as follows. First, in modern medium-level books on mathematical statistics, the decision-theoretic approach is usually presented in a rather restricted and concise manner. This practice, combined with an emphasis on estimation under the squared error loss and on testing under the zero–one loss, fails to explain why extra efforts should be made to become familiar with decision theory and to use it. Of course, dealing with selection rules requires new types of loss structures, and learning more about them leads to a better understanding of the wide range of powerful tools that decision theory has to offer. Second, permutation invariance plays an important role in selection theory. The structure of the problem of optimal permutation invariant selection rules, along with its multisample statistical model, is quite unique. Indeed, it provides a rather different setting in decision theory when compared to estimation and testing problems based on a single sample, and we wished to make those differences more readily available to our readers. In addition, as we wrote the first parts of the book, which we began in the spring of 1999, it became clear that two additional aspects of decision theory were important to us: asymptotic decision theory and the coexistence of the frequentist and Bayes approaches in decision theory. With this final realization, we settled on our main topics for the book, and they have carried us along ever since.

This book combines innovation and tradition in ways that we hope can usefully extend the line of scholarship that starts with classical monographs on decision theory by Wald (1950), Blackwell and Girshick (1954), Ferguson (1967), and DeGroot (1970) and continues with modern works by Pfanzagl and Wefelmeyer (1982, 1985), Strasser (1985), Janssen, Milbrodt, and Strasser (1985), LeCam (1986), LeCam and Yang (1990), Torgersen (1991), Bickel, Klaasen, Ritov, and Wellner (1993), Rieder (1994), and Shiryaev and Spokoiny (2000). Most of these recent publications focus primarily on fundamental structural relationships in finite and asymptotic decision theory. By contrast, we have chosen to include parts of mathematical statistics as they have been represented by Witting (1985), Lehmann (1986), Pfanzagl (1994), Witting and Müller-Funk (1995), Lehmann and Casella (1998), and Lehmann and Romano (2005). As a result, this monograph is uniquely able to synthesize otherwise disparate materials, while establishing connections between classical and modern decision theory and inviting readers to explore their interrelationships.

The importance of creating a bridge between the classical results of mathematical statistics and the modern asymptotic decision theory founded by

LeCam should not be underestimated. So far, LeCam's theory has been applied primarily to estimation and testing problems, which we now also present in the last part of Chapter 9 with treatments of selection problems. We also include new applications of this theory, which we hope demonstrate its broad and powerful applicability. The prominent monographs in that area are by Strasser (1985), LeCam (1986), Torgersen (1991), Bickel, Klaasen, Ritov, and Wellner (1993), and LeCam and Yang (2000). These are written for mathematical researchers in decision theory, and they are only partially accessible to graduate students. Representations of parts of modern decision theory, mainly applications of LAN theory to estimation and testing problems, that are accessible to graduate students can be found in the books by Behnen and Neuhaus (1989), Witting and Müller-Funk (1995), Hájek, Šidák, and Sen (1999), and Lehmann and Romano (2005). In these works, however, considerations are restricted to the asymptotic behavior of the log-likelihood under, say, a null hypothesis and local alternatives. As a consequence, the general theory of statistical models and their convergence are deliberately excluded, as are central statements of modern asymptotic decision theory. These statements provide the fundamental link between the convergence of the distributions of the likelihood ratio, the decision-theoretically motivated concept of convergence of models, and the closely related randomization criterion. They make it possible to establish the asymptotic lower Hájek–LeCam bound on the risk. The combination of this asymptotic lower bound, linearization techniques for the log-likelihood, projection techniques for the statistics, and the lemmas of LeCam constitute the backbone of modern asymptotic statistical theory. In this book we wish to present the fundamental facts and their relations to each other on an intermediate level in a form that is mathematically self-contained. This style of presentation will, we hope, enable the reader to gain deep insight into and appreciation for the structure of modern decision theory.

Another goal of this book is to provide a broad coverage of both the frequentist and the Bayes approaches in decision theory. Most existing books seem to prefer one or the other. We consider the Bayes approach to be a useful decision-theoretic framework among others, and we use it heavily throughout the book; however, we do so without extra nonmathematical philosophical justification. In this spirit we distinguish between the average risk, where the randomness of parameters is not an issue, and the Bayes risk. This distinction allows us also to treat settings with improper priors just mathematically with the average risk. Readers who are interested in contemporary presentations of Bayesian analysis, including its philosophical foundation, reasoning, and justification, are referred to the fundamental books on Bayesian analysis by Berger (1985), Bernardo and Smith (1994), Robert (2001), and Ghosh and Ramamoorthi (2003).

Chapter 1. The fundamental probabilistic concepts and technical tools are provided here. These are in the first section properties of exponential families, where we have used Brown (1986) as a guideline for our presentation.

At first glance, the importance of this class of distributions seems to be more or less due to its favorable analytical form. However, several deeper reaching characterization theorems show that, roughly speaking, finite optimal decisions are only possible for this class of distributions. The class of conjugate priors that are important for Bayes decisions, which arises in a natural way from an exponential family, is studied systematically after the Bayes framework has been introduced in Section 1.2. Tools that are used later on for Bayes estimation, testing, and selection are also prepared here.

Distances between distributions play a central role. They reflect, for example, the degree of information content in a binary model, and they explain why a decision between distributions that are farther apart is easier than a decision between distributions that are closer together. Moreover, some of the distances or transforms, (e.g., the variational and the Hellinger transforms) and their mutual relations are utilized to introduce and establish the concepts of the strong and weak convergence of statistical models. The variational and Hellinger distance, as well as the Kullback–Leibler distance, the χ^2-distance, and the Bayes error for testing hypotheses in binary models are special members of the class of v-divergences that were independently introduced by Csiszár (1963) and Ali and Silvey (1966) and constructed with the help of a convex function v. The behavior under randomization and interrelations of these functionals for different convex functions studied in Section 1.3 provides a deeper understanding of these functionals and prepares for applications in subsequent chapters. Information in Bayes models is considered next, and the chapter concludes with an introduction to $\mathbb{L}_2$-differentiability, where we have used Witting (1985) as a guideline for our presentation.

Chapter 2. The central topic is the Neyman–Pearson lemma and its extensions. Links between Neyman–Pearson, minimax, and Bayes tests are discussed and studied in detail. After a consideration of statistical models with stochastic ordering, especially with a monotone likelihood ratio, which include exponential families, Neyman–Pearson's lemma is extended to tests for composite one-sided hypotheses.

Chapter 3. An introduction to the general framework of decision theory is given, followed by a discussion of its components. The concept of convergence of decisions and the sequentially weak compactness of the set of all decisions for a given model are central topics. Here and at several other places of the book, we restrict ourselves to compact decision spaces and dominated models. This practice helps keep technical tools at the graduate level, and it usefully restricts references to results in other literature.

Special properties of the risk as a function of the parameter as well as of the decision are studied to prepare for theorems of the existence of Bayes and minimax decisions. Furthermore, the interrelations between Bayes and minimax decisions are studied in preparation of proofs of minimaxity of estimators and tests later on that are based on Bayes properties and a constant risk. Γ-minimax decisions, which are analogues to minimax decisions in the

Bayes approach, are also briefly considered in Section 3.6, and the chapter concludes with special versions of the minimax theorem and the complete class theorem. For readers interested in further results, references are made to the fundamental monographs by Strasser (1985) and LeCam (1986).

Chapter 4. The chapter begins with examples in which randomizations of models appear in a natural way. The concept of ε-deficiency due to LeCam (1964), which is a comparison of the risk function "up to ε", is essential for the approximation and convergence of models and takes the center stage in this chapter. Another fundamental result is the randomization theorem of decision theory. It shows that the decision-theoretic concept of ε-deficiency is identical with the variational distance between one model and a suitable randomization of the other model. A transition to standard models gives the statement that finite models are uniquely determined by their standard distributions and the Hellinger tranforms. The characterization of the ε-deficiency via Bayes risks leads to the concept of standard decision problems for which the associated risk is just a special v-divergence. This is the concave function criterion of decision theory, and it connects concepts from information and decision theory.

In the second part of the chapter, sufficient statistics are characterized by the fact that the induced model is equivalent to the original model. The v-divergences are used to give for the sufficiency an information-theoretic characterization due to Csiszár (1963), the test-theoretic characterization due to Pfanzagl (1974), and the well-known factorization criterion by Neyman. A discussion of the different concepts of sufficiency such as pairwise sufficiency, Blackwell sufficiency, and Bayes sufficiency is included. A brief discussion of ancillarity, which includes Basu's theorem, concludes the chapter.

Chapter 5. The treatment of the reduction by invariance is kept concise by mainly considering the groups of permutations, location-scale transforms, and rotations. Whereas permutation invariance is especially relevant for selection rules, the other groups are utilized to prove the Hunt–Stein theorem on the minimaxity of best invariant tests. Hereby the existence of the Haar measure can be established directly in a simple manner without having recourse to further literature. The connection between best equivariant estimators and minimax estimators is provided by the Girshick–Savage theorem. With the conclusion of this chapter all tools from finite decision theory that are necessary for our purposes have been collected.

Chapter 6. The previous results on ε-deficiency and the randomization theorem are used to develop a theory of convergence of models within our fixed framework. Asymptotically normal models play a central role. Whereas the term *model* is standard in mathematical statistics, the term *experiment* is more common in modern decision theory. As both concepts are essentially the same (see Lehmann and Romano (2005) p. 550), we use the term *model* throughout the book. The transition to standard models makes it possible to get for finite models the well-known bounds on the ε-deficiency in terms of the Dudley metric of standard distributions, which leads to the characterization of

the convergence of finite models in terms of the distributions of the likelihood ratios. For binary models the concepts of contiguity and entire separation are introduced through the accumulation points of a sequence of models. As in Jacod and Shiryaev (1987, 2002) and Liese (1986), we use Hellinger integrals to get the results on the contiguity and the entire separation of sequences of binary models, especially the results of Oosterhoff and van Zwet (1979) for triangular arrays of independent models. In the study of the asymptotic normality of double sequences of binary models, we follow the ideas of LeCam and Yang (1990).

After the introduction and brief discussion of Gaussian models the LAN- and ULAN-properties are introduced and established for localized sequences of differentiable models. From the start, after Witting and Müller-Funk (1995) and Rieder (1994), regression coefficients that satisfy the Noether condition are used. Special cases then are the row-wise i.i.d. case, the two-sample problem, and regression models with deterministic covariables. Suitable versions of the third lemma of LeCam are given. These results allow us to study the risks of sequences of decisions in a shrinking sequence of the localization point of the models, providing a comparison of the efficiency of different sequences of decisions.

In the remainder of the chapter, the lower Hájek–LeCam bound is derived. To avoid advanced techniques from topology, the bound is established here only for compact decision spaces and dominated limit models. This proves sufficient for our purposes, as the models considered here are nearly always parametric models. The lower Hájek–LeCam bound makes it possible to break up the proof of asymptotic optimality of estimators, tests, and selection rules into separate steps. The first step consists of finding in the asymptotic Gaussian model the optimal solution, which depends only on the sufficient central variable. By replacing the central variable with the central sequence a sequence of decisions is obtained. Under additional regularity assumptions the convergence of the risks to the lower Hájek–LeCam follows, and this in turn guarantees the optimality of the sequence of decisions.

Chapter 7. The chapter on parameter estimation begins with the Cramér–Rao inequality and the result, which has been proved by various authors under different regularity assumptions: namely that equality only holds for exponential families. This result corroborates the importance of exponential families for statistical analyses under finite sample sizes and it distinguishes a need for asymptotic considerations. Classical results on UMVU estimators, selected topics on Bayes estimators, and considerations regarding the admissibility of estimators conclude the first part of this chapter.

The second part is devoted to the study of the asymptotic properties of estimators of parameters. For all asymptotic considerations it is mandatory to deal first with the question of the consistency of estimators. Only for estimators with this property can classical and modern linearization techniques be utilized. From a variety of possible approaches to consistency we have

chosen the concept of M-estimators, and we follow here to some extent the presentation in Pfanzagl (1994). Besides a treatment of the consistency of M-estimators and the MLEs, and a discussion of the existence of MLEs in exponential families, we study location and regression models. Techniques from convex analysis, due to Hjort and Pollard (1993), allow us to verify consistency without assumptions regarding compactness for convex criterion functions. The part on consistency is completed with the consistency in Bayes models. In giving the fundamental results of Doob (1949) and Schwartz (1965) we follow Ghosh and Ramamoorthi (2003). One way of proving the asymptotic normality of M-estimators is based on the classical Taylor expansion. However, for the treatment of regression models with not necessarily differentiable criterion functions it is preferable to follow linearization techniques for convex criterion functions based on Hjort and Pollard (1993). Doing so avoids conditions regarding differentiability. The necessity of taking the second way arises in $\mathbb{L}_1$-regression and more generally in quantile regressions, as they are represented in Jurečková and Sen (1996). The asymptotic normality of the posterior distribution (i.e., the Bernstein–von Mises theorem) is established and used to prove the asymptotic normality of the Bayes estimator.

The last section of this chapter deals with the asymptotic optimality of the MLE. The result by Bahadur (1964) on the majorization of the covariance matrix of the limit distribution of an asymptotically normal estimator over the inverse of Fisher's information matrix is presented. Then the estimation problem is treated systematically as a decision problem, and the lower bound on the risks is derived under different conditions by utilizing the general results from Chapter 6. This is done in the finite-dimensional case for the asymptotically median unbiased estimators. In the multivariate case, an asymptotic minimax bound is derived. It is shown that in each case, under weak assumptions the MLE achieves the respective lower bound. With these main theorems in asymptotic estimation theory this chapter is completed.

Chapter 8. At the beginning uniformly best unbiased level α tests for two-sided hypotheses in one-parameter exponential families are characterized. Then there follows a section on testing linear hypotheses in multivariate normal distributions with a common known covariance matrix. These results constitute, from an asymptotic point of view, the solution of the decision problem in the limit model. Uniformly best unbiased level α tests in d-parameter exponential families, which are conditional tests, are derived next. Selected topics on uniformly best invariant level α tests and Bayes tests conclude the first part of this chapter.

The second part is devoted to the study of the asymptotic properties of tests. It begins with the study of exponential rates of error probabilities in binary models, which leads to the theorems of Stein and Chernoff. The major treatment of asymptotic tests starts with a problem that is of importance of its own: the central question about the linearizations of statistics. Whereas in the area of parameter estimation such linearizations are the result of the

linearization of equations, supporting tools of this type are not available for tests. For the latter, the projection techniques due to Hájek are fundamental. This has been used already for U-statistics in the special case of Hoeffding. The usefulness of these projection techniques is demonstrated on U-statistics and rank statistics, which serve as preparation for the results on the local asymptotic optimality of linear rank tests. Projection techniques are also used to study statistics that include estimated nuisance parameters.

The results on the linearization of statistics are used to establish the asymptotic normality of the test statistics under the null hypothesis. A combination of the asymptotic upper Hájek–LeCam bound for the power with the third lemma of LeCam allows the characterization of locally asymptotically most powerful tests and the calculation of the relative efficiency of given tests. For one-dimensional and multivariate parameters of interests, in models with or without nuisance parameters, we characterize the locally asymptotically optimal tests. In particular, we study Neyman's score test, the likelihood ratio tests, and tests that are based on the MLE known as Wald tests. The asymptotic relative efficiency of selected rank tests, especially for the two-sample problem, is determined by investigating the local asymptotic power along parametric curves in the space of the distributions. For given rank tests, the parametric models are determined for which these rank tests are locally asymptotically best.

Chapter 9. Selection rules are presented here within the decision theoretic framework of the book. The goal is to select a best, or several of the best, of k independent populations. The foundation of finite sample size selection rules goes back to Paulson (1949, 1952), Bahadur and Goodman (1952), Bechhofer (1954), Bechhofer, Dunnett, and Sobel (1954), Gupta (1956, 1965), Lehmann (1957a,b, 1961, 1963, 1966), and Eaton (1967a,b). The first research monographs were written by Bechhofer, Kiefer, and Sobel (1968), Gibbons, Olkin, and Sobel (1977), and Gupta and Panchapakesan (1979).

After an introduction of the selection models, optimal point selection rules are derived for parametric and especially for exponential families. For equal sample sizes the fundamental Bahadur–Goodman–Lehmann–Eaton theorem states that the natural selection rule is the uniformly best permutation invariant decision. For unequal sample sizes the situation changes dramatically and the natural selection rule loses many of its qualities (see Gupta and Miescke (1988)). Bayes selection rules in explicit form are not always readily available. For exponential families, conjugate priors can be chosen such that the posterior distributions are balanced and provide Bayes solutions of a simple form. Combining selection with the estimation of the parameter of the selected population is also considered. The next section deals with subset selections and especially with Gupta's subset selection rule. Γ-minimax selections are also considered here. Section 9.3 deals with multistage selection rules that improve the efficiency by combining the approaches of the previous two sections

(see, e.g., Miescke (1984a, 1999)). Selected results, including Bayes designs for stagewise sampling allocations, are presented in detail.

The second part of the chapter is on asymptotic properties of selection rules, and it starts with the exponential rates of the error probabilities of selection rules from Liese and Miescke (1999a). These results are related to results of Chernoff (1952, 1956) and Krafft and Puri (1974). Then localized parametric models are considered. It is shown that under equal sample sizes the natural selection rule based on the central sequence is both locally asymptotically uniformly best in the class of all permutation invariant selection rules and locally asymptotically minimax in terms of the pointwise comparison of the asymptotic risks. Because the statistics used by the selection rules have a specific difference structure, which is similar to the situation of two-sample problems, the localization point that appears in the central sequence can be replaced by an estimator without changing the asymptotic efficiency. The same holds true for additional nuisance parameters. In the nonparametric selection model we study selection rules that are based on rank statistics. Here we use results that have been prepared previously for nonparametric tests.

There are a number of people and institutions that we would like to thank for supporting this book project. Several rounds of reviews over the past three years have given us immeasurable help getting the book into shape, and we are deeply indebted to all of the experts who were willing to review our material and provide critical comments and suggestions. We are very grateful to the Mathematical Research Institute at Oberwolfach for letting us stay and work within its RIP program for two weeks in both 2004 and 2005. The support we have received from Springer Verlag and the guidance we have received from John Kimmel and his technical staff have greatly facilitated our work, and we are especially appreciative. We also thank the colleagues in our departments who have contributed to countless discussions throughout the progress of the book. Their input as well as their understanding of our long preoccupation with this project are very much appreciated. Additionally, thanks are due to our departments and universities for the time and working space that they have provided us. We thank Peter Dencker and Jin Tan for proofreading parts of the book and Jenn Fishman for helping us with the revision of the preface.

Our special thanks go to Ingo Steinke, who proofread several versions of the book, pointed out many inaccurate details, and provided valuable suggestions for improving the book's overall layout. His continuous interest and help in this project is highly appreciated.

Finally, we would like to say some words in memory of Shanti S. Gupta, who passed away in 2002. His inspiration, support, and encouragement have deeply affected our lives and, in particular, our work on this book.

Rostock and Chicago, *Friedrich Liese*
October 2007 *Klaus-J. Miescke*

Contents

1

Statistical Models

The starting point of all statistical inference is the observation of data that are subject to unavoidable random errors. The intention is to draw conclusions from the data in such a way that the information that is contained in the data is exploited as much as possible. For this purpose we need a mathematical model that explains the fluctuation of the observations from measurement to measurement, then a mathematical frame for possible conclusions, and finally a tool for the assessment of the quality of concrete conclusions. Although usually error-free conclusions from disturbed data cannot be drawn, we can improve, or even optimize, the inference by utilizing our knowledge of the probabilities of the random events that are relevant for the statistical problem at hand.

The basic object is a suitably chosen space $\mathcal{X}$ in which all concrete measurements can be observed. Following standard practice in probability theory let there be given a σ-algebra $\mathfrak{A}$ of subsets of $\mathcal{X}$ so that $\mathfrak{A}$ contains all subsets of $\mathcal{X}$ that are relevant for the problem. The pair $(\mathcal{X}, \mathfrak{A})$ is called the *sample space*. If $\mathcal{X}$ is a metric space, then we use the Borel sets as the σ-algebra $\mathfrak{A}$. On the other hand, if $\mathcal{X}$ is finite or countably infinite, then we use the power set $\mathfrak{P}(\mathcal{X})$ for $\mathfrak{A}$.

To explain the fluctuation of the observations we assume that each observation $x \in \mathcal{X}$ is the realization of a random variable X with values in $\mathcal{X}$ that is defined on some underlying abstract *probability space* $(\Omega, \mathfrak{F}, \mathbb{P})$, where $\mathbb{P}$ is a probability measure on $(\Omega, \mathfrak{F})$. By definition, such a random variable is a mapping $X : \Omega \to \mathcal{X}$ that is $\mathfrak{F}$-$\mathfrak{A}$ measurable, i.e., $X^{-1}(A) \in \mathfrak{F}$, $A \in \mathfrak{A}$, where $X^{-1}(A) = \{\omega : X(\omega) \in A, \omega \in \Omega\}$. To indicate that X is measurable we use the notation $X : \Omega \to_m \mathcal{X}$.

To be able to work on concrete problems a link to a family of concrete probability spaces, say $(\mathcal{X}, \mathfrak{A}, P_\theta)$, $\theta \in \Delta$, has to be established by means of possible distributions P_θ of X at $\theta \in \Delta$ that include the true but unknown distribution of X. This leads to the concept of a *statistical model*. The first step toward a statistical model is to choose a suitable family $(P_\theta)_{\theta \in \Delta}$ of distributions of X on $(\mathcal{X}, \mathfrak{A})$. This can be a difficult and challenging task, depending

F. Liese, K.-J. Miescke, *Statistical Decision Theory*,
DOI: 10.1007/978-0-387-73194-0_1,

on the experimental situation. The choice has to be made based on the initial information that is available about the random behavior of X in the experiment. To be mathematically consistent we assume that there is a family of probability measures $(\mathbb{P}_\theta)_{\theta\in\Delta}$ on $(\Omega,\mathfrak{F})$ such that for every $\theta\in\Delta$ the distribution of X under $\mathbb{P}_\theta$ is given by $P_\theta=\mathbb{P}_\theta\circ X^{-1}$; i.e., $P_\theta(A)=\mathbb{P}_\theta(X\in A)$, $A\in\mathfrak{A}$. By combining the sample space with the set of possible distributions of X we arrive at the statistical model

$$\mathcal{M}=(\mathcal{X},\mathfrak{A},(P_\theta)_{\theta\in\Delta}). \tag{1.1}$$

If the *parameter set* Δ is finite, then we call $\mathcal{M}$ a *finite model*. The simplest models are *binary models* where Δ consists only of two elements.

1.1 Exponential Families

Many of the frequently used parametric families of distributions $(P_\theta)_{\theta\in\Delta}$ in a statistical model $\mathcal{M}=(\mathcal{X},\mathfrak{A},(P_\theta)_{\theta\in\Delta})$ are special cases of exponential families. Examples are the normal, binomial, Poisson, beta, and gamma families. Because all of these families share properties that are typical for an exponential family, it is natural and proves useful to study first this important general statistical model, and to collect analytical properties that are used throughout this book.

We are following here the tradition set by Lehmann (1959, 1983) in his classical books on testing and estimation, and continued in their respective second editions: Lehmann (1986), Lehmann and Casella (1998), and Lehmann and Romano (2005). More general treatments of exponential families are provided in Barndorff-Nielsen (1978) and Brown (1986). We also refer to Hoffmann-Jørgensen (1994), Johansen (1979), and Küchler and Sørensen (1997).

Let $(\mathcal{X},\mathfrak{A})$ be a given measurable space and $T:\mathcal{X}\to_m\mathbb{R}^d$ be a *statistic*. For any $\boldsymbol{\mu}\in\mathcal{M}^\sigma(\mathfrak{A})$, we take

$$\Delta=\{\theta:\int\exp\{\langle\theta,T\rangle\}d\boldsymbol{\mu}<\infty\}\subseteq\mathbb{R}^d,\quad\text{and} \tag{1.2}$$

$$K(\theta)=\ln(\int\exp\{\langle\theta,T\rangle\}d\boldsymbol{\mu}),\quad\theta\in\Delta, \tag{1.3}$$

where $\langle\theta,T\rangle=\theta^TT=\sum_{i=1}^d\theta_iT_i$ is the Euclidean scalar product of the vectors $\theta=(\theta_1,...,\theta_d)^T$ and $T=(T_1,...,T_d)^T$. Given $0<\alpha<1$, we set $p=1/\alpha$ and $q=1/(1-\alpha)$. Then $1/p+1/q=1$ and by Hölder's inequality (see Lemma A.13) it holds for $\theta_1,\theta_2\in\Delta$,

$$\begin{aligned}\exp\{K(\alpha\theta_1+(1-\alpha)\theta_2)\} &= \int \exp\{\langle \alpha\theta_1 + (1-\alpha)\theta_2, T\rangle\} d\boldsymbol{\mu} \qquad (1.4)\\ &= \int \exp\{\langle \alpha\theta_1, T\rangle\} \exp\{\langle (1-\alpha)\theta_2, T\rangle\} d\boldsymbol{\mu}\\ &\le (\int \exp\{\langle \theta_1, T\rangle\} d\boldsymbol{\mu})^{\alpha} (\int \exp\{\langle \theta_2, T\rangle\} d\boldsymbol{\mu})^{1-\alpha}\\ &= \exp\{\alpha K(\theta_1) + (1-\alpha)K(\theta_2)\}.\end{aligned}$$

This means that the set Δ in (1.2) is a convex set, and that the function K in (1.3) is convex. For every $\theta \in \Delta$,

$$P_\theta(A) = \int_A \exp\{\langle \theta, T\rangle - K(\theta)\} d\boldsymbol{\mu},\ A \in \mathfrak{A}, \qquad (1.5)$$

is a probability measure on $(\mathcal{X}, \mathfrak{A})$, and the family of distributions $(P_\theta)_{\theta\in\Delta}$ is called an *exponential family*. We denote by

$$f_\theta(x) := \frac{dP_\theta}{d\boldsymbol{\mu}}(x) = \exp\{\langle \theta, T(x)\rangle - K(\theta)\}, \quad x \in \mathcal{X}, \qquad (1.6)$$

the density of P_θ with respect to $\boldsymbol{\mu}$, $\theta \in \Delta$.

It should be noted that, in general, the parameter set Δ is neither closed nor open. An exponential family $(P_\theta)_{\theta\in\Delta}$ is called *regular* if $\Delta = \Delta^0$, where here and in the sequel Δ^0 denotes the interior of Δ.

Throughout the book, whenever an exponential family is considered, the following two assumptions are made to make sure that the dimensions of $\mathbb{R}^d$ and Δ can not be reduced.

(A1) The statistics $T_1, ..., T_d$ are linearly independent in the sense that for $a_0, a_1, ..., a_d \in \mathbb{R}$, the relation $a_1T_1 + \cdots + a_dT_d = a_0$, $\boldsymbol{\mu}$-a.e., implies that $a_i = 0$, $i = 0, 1, ..., d$.

(A2) The interior Δ^0 of Δ is nonempty.

If the condition (A1) is fulfilled, then the parameter θ is *identifiable*; that is, $P_{\theta_1} = P_{\theta_2}$ implies $\theta_1 = \theta_2$. If not already achieved from the very beginning, the technical tools of reparametrization and a suitable choice of the measure $\boldsymbol{\mu}$ are available for this purpose.

Definition 1.1. *Under the assumptions (A1) and (A2) made on T and Δ, respectively, the family of distributions $(P_\theta)_{\theta\in\Delta}$ given by* (1.5) *is called a d-parameter exponential family in natural form, with natural parameter θ and generating statistic T. The statistical model*

$$\mathcal{M}_{ne} = (\mathcal{X}, \mathfrak{A}, (P_\theta)_{\theta\in\Delta}) \qquad (1.7)$$

with $(P_\theta)_{\theta\in\Delta}$ from (1.5) *is called a natural exponential model. It is called regular if Δ is open.*

An exponential family in natural form is also called an *exponential family in canonical form* in the literature.

Problem 1.2.* The representation of an exponential family in natural form by (1.5) is not unique in the triplet $(T, \theta, \boldsymbol{\mu})$.

The ambiguity pointed out in the above problem is often utilized to find representations that are better adapted to the problem under consideration.

As the density in (1.5) is positive we see that the distributions from the exponential family are measure-theoretically equivalent to $\boldsymbol{\mu}$; that is, $P_\theta(B) = 0$ if and only if $\boldsymbol{\mu}(B) = 0$, or in short $\boldsymbol{\mu} \ll\gg P_\theta$. This implies

$$P_\theta \ll\gg P_{\theta_0}, \quad \theta_0, \theta \in \Delta, \tag{1.8}$$

and that the density of P_θ with respect to P_{θ_0} is given by

$$\frac{dP_\theta}{dP_{\theta_0}} = \exp\{\langle \theta - \theta_0, T\rangle - K(\theta) + K(\theta_0)\}, \quad \theta_0, \theta \in \Delta.$$

For $d = 1$ the condition (A1) only means that the statistic T is not $\boldsymbol{\mu}$-a.e. constant and therefore in view of (1.8) T is not P_θ-a.s. constant. For $d > 1$ the condition (A1) excludes the cases where P_θ-a.s. the statistic T takes on values in a lower-dimensional subspace. We show later that $\mathsf{E}_\theta \|T\|^2 < \infty$, $\theta \in \Delta$. For such a random vector the fact that only values from a subspace are attained can be characterized with the help of the covariance matrix.

Problem 1.3.* Let $Y_1, ..., Y_d$ be random variables with finite second moments. There exist $a_0, a_1, ..., a_d \in \mathbb{R}$ with $\sum_{i=1}^d a_i^2 > 0$ and $a_1 Y_1 + \cdots + a_d Y_d = a_0$, $\mathbb{P}$-a.s., if and only if the covariance matrix of $(Y_1, ..., Y_d)$ is singular.

For some purposes it proves convenient to study the family of induced distributions $Q_\theta = P_\theta \circ T^{-1}$. The statistical model

$$\mathcal{M}_{re} = (\mathbb{R}^d, \mathfrak{B}_d, (Q_\theta)_{\theta\in\Delta}), \tag{1.9}$$

is called the *reduced model* or the *model in minimal form*. For every $B \in \mathfrak{B}_d$ and $\boldsymbol{\nu} = \boldsymbol{\mu} \circ T^{-1}$,

$$Q_\theta(B) = \int I_B(T) \exp\{\langle \theta, T\rangle - K(\theta)\} d\boldsymbol{\mu} = \int_B \exp\{\langle \theta, t\rangle - K(\theta)\} \boldsymbol{\nu}(dt),$$

so that

$$g_\theta(t) := \frac{dQ_\theta}{d\boldsymbol{\nu}}(t) = \exp\{\langle \theta, t\rangle - K(\theta)\}, \quad t \in \mathbb{R}^d. \tag{1.10}$$

When passing from the natural form to the reduced form we changed the sample space with the consequence that the new generating statistic (i.e., the identical mapping) is very simple. Later we show that the two models, the natural model and the reduced model, are identical from the decision-theoretic point of view.

It is an important property of exponential families that the distributions of a sample of size n form again an exponential family where the new generating statistic is the sum. The following proposition presents the precise statement which is a consequence of the fact that the density of a product measure with respect to another product measure is simply the product of the individual densities; see Proposition A.29.

Proposition 1.4. *Let $(P_\theta)_{\theta\in\Delta}$ be a natural exponential family with respect to $\boldsymbol{\mu}$. Then $(P_\theta^{\otimes n})_{\theta\in\Delta} \subseteq \mathcal{P}(\mathfrak{A}^{\otimes n})$ is a natural exponential family with respect to $\boldsymbol{\mu}^{\otimes n}$ with generating statistic $T_{\oplus n}(x_1,...,x_n) := \sum_{i=1}^n T(x_i)$ and it holds that*

$$\frac{dP_\theta^{\otimes n}}{d\boldsymbol{\mu}^{\otimes n}} = \exp\{\langle\theta, T_{\oplus n}\rangle - nK(\theta)\}.$$

If X and Y are independent random vectors with distributions P and Q, respectively, then the distribution of $X+Y$ is given by the convolution of the two distributions P and Q, defined by

$$(P * Q)(B) := \int P(B-x)Q(dx), \quad B \in \mathfrak{B}_d.$$

According to (1.9) the reduced version of $P_\theta^{\otimes n}$ is given by $Q_{n,\theta} := P_\theta^{\otimes n} \circ T_{\oplus n}^{-1}$. As $T_{\oplus n}$ is the sum of n independent identically distributed (i.i.d.) random vectors we see that the reduced model is given by

$$Q_{n,\theta} = \mathcal{L}(T_{\oplus n}|P_\theta^{\otimes n}) = (P_\theta \circ T^{-1})^{*n},$$

where $*n$ denotes the n-fold convolution.

For practical purposes we may also change the parameter set. Such a reparametrization can often be made to get new parameters that allow for a better statistical interpretation. Let $\Lambda \subseteq \mathbb{R}^d$ and $\kappa : \Lambda \to \Delta$ be a mapping. Then (1.5) can be reparametrized to

$$P_{pe,\eta}(A) := P_{\kappa(\eta)}(A) = \int_A \exp\{\langle\kappa(\eta), T\rangle - K(\kappa(\eta))\}d\boldsymbol{\mu}, \quad A \in \mathfrak{A}, \tag{1.11}$$
$$h_\eta(x) := \frac{dP_{pe,\eta}}{d\boldsymbol{\mu}}(x) = \exp\{\langle\kappa(\eta), T(x)\rangle - K(\kappa(\eta))\}, \quad x \in \mathcal{X},$$

where $\eta \in \Lambda$. The statistical model

$$\mathcal{M}_{pe} = (\mathcal{X}, \mathfrak{A}, (P_{pe,\eta})_{\eta\in\Lambda}), \tag{1.12}$$

with $(P_{pe,\eta})_{\eta\in\Lambda}$ from (1.11), is called a *reparametrized exponential model*. Whenever the representation (1.12) is used, we assume without loss of generality that the mapping $\kappa : \Lambda \to \Delta$ is a one-to-one mapping of Λ into Δ. This guarantees that for any two parameter points $\eta_1, \eta_2 \in \Lambda$, $P_{pe,\eta_1} = P_{pe,\eta_2}$ implies $\eta_1 = \eta_2$. In this case, the parameter η in the family $(P_{pe,\eta})_{\eta\in\Lambda}$ is identifiable. Moreover, we use $\gamma = \kappa^{-1}$ in the sequel. A concrete statistical model

usually is introduced by specifying $(P_{pe,\eta})_{\eta\in\Lambda}$, where the parameter η admits a direct statistical interpretation.

In the following examples, we look at some common parametric families of distributions and represent them as exponential families. As the natural parameter is not necessarily the parameter that admits a statistical interpretation we often introduce another more meaningful parameter.

Here and in the sequel, whenever an at most countable sample space $\mathcal{X}$ appears we use the power set $\mathfrak{P}(\mathcal{X})$, i.e., the system of all subsets of $\mathcal{X}$, as σ-algebra $\mathfrak{A}$ in our statistical model. Unless explicitly mentioned otherwise, we use the counting measure as the dominating measure so that we have only to deal with the probability mass function (p.m.f.), $f(x) := P(\{x\})$, $x \in \mathcal{X}$, which is the density of P with respect to the counting measure. We set

$$\begin{aligned}
\mathbb{S}_{d-1} &= \{(p_1,...,p_{d-1}) : p_i > 0,\ i = 1,...,d-1,\ \textstyle\sum_{j=1}^{d-1} p_j < 1\},\\
\mathbf{S}_d^o &= \{(p_1,...,p_d) : p_i > 0,\ i = 1,...,d,\ \textstyle\sum_{j=1}^{d} p_j = 1\},\\
\mathbf{S}_d^c &= \{(p_1,...,p_d) : p_i \geq 0,\ i = 1,...,d,\ \textstyle\sum_{j=1}^{d} p_j = 1\}.
\end{aligned} \tag{1.13}$$

Example 1.5. Let $X_1,...,X_n$ be a sample of i.i.d. observations from an experiment with d possible outcomes that have probabilities p_i, $i = 1,...,d$. The sample space is $\mathcal{X} = \{(\varepsilon_1,...,\varepsilon_n) : \varepsilon_i \in \{1,...,d\},\ i = 1,...,n\}$ and it holds

$$\begin{aligned}
\mathbb{P}(X_1 = \varepsilon_1,...,X_n = \varepsilon_n) &= \textstyle\prod_{i=1}^{n} p_{\varepsilon_i}\\
&= \exp\{\textstyle\sum_{j=1}^{d} T_j(x)\ln p_j\}, \quad x = (\varepsilon_1,...,\varepsilon_n) \in \mathcal{X}, \quad \text{where}\\
T_j(x) &= |\{i : \varepsilon_i = j,\ \ i = 1,...,n\}|
\end{aligned}$$

is the number of observations with outcome j, $j = 1,...,d$. As $\sum_{j=1}^{d} p_j = 1$ the assumption (A2) is not met. However, by a reduction to $d-1$ parameters we can get an exponential family (1.5) in natural form. Put for $i = 1,...,d-1$,

$$\begin{aligned}
&\theta_i = \kappa_i(p) := \ln(p_i/p_d), \quad p_i = \gamma_i(\theta) = \exp\{\theta_i\}(1 + \textstyle\sum_{j=1}^{d-1}\exp\{\theta_j\})^{-1},\\
&p_d = 1 - \textstyle\sum_{i=1}^{d-1} p_i, \quad p_d = \gamma_d(\theta) = 1 - \textstyle\sum_{i=1}^{d-1}\gamma_i(\theta),\\
&T = (T_1,...,T_{d-1}), \quad K(\theta) = n\ln(1 + \textstyle\sum_{j=1}^{d-1}\exp\{\theta_j\}),\\
&\theta = (\theta_1,...,\theta_{d-1}) \in \Delta = \mathbb{R}^{d-1}, \quad p = (p_1,...,p_d) \in \mathbf{S}_d^o.
\end{aligned}$$

As $\sum_{i=1}^{d} k_i \ln p_i = \sum_{i=1}^{d} k_i\theta_i$ we see that $P_\theta = \mathcal{L}(X_1,...,X_n)$ has the p.m.f.

$$\begin{aligned}
f_\theta(x) &= \exp\{\textstyle\sum_{i=1}^{d} T_i(x)\ln p_i\}\\
&= \exp\{\textstyle\sum_{i=1}^{d-1} T_i(x)\ln p_i + (n - \sum_{i=1}^{d-1} T_i(x))\ln(1 - \sum_{i=1}^{d-1} p_i)\}\\
&= \exp\{\textstyle\sum_{i=1}^{d-1}\theta_i T_i(x) - K(\theta)\}.
\end{aligned}$$

With $\boldsymbol{\mu}$ being the counting measure we see that $(P_\theta)_{\theta\in\Delta}$is a regular $(d-1)$ parameter exponential family with natural parameter $\theta \in \mathbb{R}^{d-1}$ that satisfies (A1) and (A2). The distributions in the reduced model are then

$$P_\theta \circ T^{-1} = \mathsf{M}(n, \gamma(\theta)), \quad \theta \in \mathbb{R}^{d-1},$$

where $\mathsf{M}(n,p)$ denotes the multinomial distribution with parameters n and $p = (p_1, ..., p_d) \in \mathbf{S}_d^o$.

Problem 1.6. Verify the statements in the previous example regarding (A1) and (A2).

Problem 1.7. Let $X_1, ..., X_n$ be i.i.d. Bernoulli variables with success probability $p \in (0,1)$. Then the joint distribution on $\mathcal{X} = \{0,1\}^n$ is given by $((1-p)\boldsymbol{\delta}_0 + p\boldsymbol{\delta}_1)^{\otimes n}$, where $\boldsymbol{\delta}_a$ is the δ-distribution that is concentrated at point a. Set

$$\theta = \kappa(p) := \ln(p/(1-p)), \quad p = \gamma(\theta) := \frac{\exp\{\theta\}}{1+\exp\{\theta\}},$$
$$K(\theta) = n\ln(1+e^\theta), \quad \Delta = \mathbb{R},$$
$$T(x) = \sum\nolimits_{i=1}^n x_i, \quad x = (x_1, ..., x_n) \in \{0,1\}^n.$$

Then the family of distributions $(P_\theta)_{\theta\in\Delta} = ((1-\gamma(\theta))\boldsymbol{\delta}_0 + \gamma(\theta)\boldsymbol{\delta}_1)^{\otimes n}$ has the p.m.f. $f_\theta = \exp\{\theta T - K(\theta)\}$ and is thus a one-parameter exponential family with natural parameter θ and generating statistic T. The distributions in the reduced model are $P_\theta \circ T^{-1} = \mathsf{B}(n, \gamma(\theta))$, $\theta \in \mathbb{R}$.

Problem 1.8.* Sometimes, the parameter set $(0,1)$ of the binomial distribution $\mathsf{B}(n,p)$ is extended by putting $\mathsf{B}(n,0) = \boldsymbol{\delta}_0$ and $\mathsf{B}(n,1) = \boldsymbol{\delta}_n$. Show that the extended family $\mathsf{B}(n,p)$, $p \in [0,1]$, cannot be represented as an exponential family.

Problem 1.9.* The family of Poisson distributions $(\mathsf{Po}(\lambda))_{\lambda>0}$ with p.m.f.

$$\mathsf{po}_\lambda(k) = \frac{\lambda^k}{k!}\exp\{-\lambda\}, \quad k \in \mathbb{N}, \quad \lambda > 0,$$

can be represented as a one-parameter exponential family in natural form.

Example 1.10. The exponential families in Example 1.5 and in the Problems 1.7 and 1.9 are regular, i.e., their natural parameter sets are open. This property is often met, but there is an important exponential family that does not share this property. Let $W(t)$, $t > 0$, be a standard Wiener process and ν and σ be fixed positive constants . For $a > 0$ we denote by $T_a = \inf\{t : \nu t + \sigma W(t) \geq a\}$ the first passage time at which the process $\nu t + \sigma W(t)$ crosses the level a. It can be shown (see Seshadri (1993)) that T_a is finite with probability one and that it has a distribution, called the inverse Gaussian distribution $\mathsf{Gi}(\lambda, m)$, that has the Lebesgue density

$$\mathsf{gi}_{\lambda,m}(x) = \sqrt{\frac{\lambda}{2\pi x^3}}\exp\{-\frac{\lambda}{2m^2}\frac{(x-m)^2}{x}\}I_{(0,\infty)}(x), \tag{1.14}$$

where $\lambda = (a/\sigma)^2$ and $m = a/\nu$. Letting $m \to \infty$ we get as the density of the first passage time of the standard Wiener process

$$\mathsf{gi}_{\lambda,\infty}(x) = \sqrt{\frac{\lambda}{2\pi x^3}}\exp\{-\frac{\lambda}{2x}\}I_{(0,\infty)}(x). \tag{1.15}$$

To present the densities $\mathsf{gi}_{\lambda,m}$ in standard exponential form we set $\mathcal{X} = (0,\infty)$ and introduce the measure $\boldsymbol{\mu}$ by $\boldsymbol{\mu}(dx) = (2\pi x^3)^{-1/2}\boldsymbol{\lambda}(dx)$ on $\mathfrak{B}_{(0,\infty)}$. If $T_1(x) = x$, $T_2(x) = 1/x$,

$$(\theta_1,\theta_2) = (-\frac{\lambda}{2m^2}, -\frac{\lambda}{2}), \quad K(\theta_1,\theta_2) = 2\sqrt{\theta_1\theta_2} - \frac{1}{2}\ln(-2\theta_2),$$

then

$$\frac{d\mathsf{Gi}(\lambda,m)}{d\boldsymbol{\mu}}(x) = \exp\{\theta_1 x + \theta_2\frac{1}{x} - K(\theta_1,\theta_2)\}.$$

The natural parameter set is $\Delta = (-\infty,0]\times(-\infty,0)$ which is not open. This set corresponds to the set $(0,\infty)\times(0,\infty]$ in the original parametrization.

Normal distributions are exponential families. The two-parameter case is studied in the next example. The one-parameter cases, where either the variance or the mean is known, are considered in Lemma 1.37 and Example 1.38, respectively.

Example 1.11. Let X be an observation from a normal distribution $\mathsf{N}(\mu,\sigma^2)$, where $(\mu,\sigma^2)\in\Lambda = \mathbb{R}\times(0,\infty)$ is unknown. The density φ_{μ,σ^2} of the distribution $\mathsf{N}(\mu,\sigma^2)$ with respect to the Lebesgue measure $\boldsymbol{\lambda}$ on $\mathbb{R}$ is

$$\begin{aligned}\varphi_{\mu,\sigma^2}(x) &= (2\pi\sigma^2)^{-1/2}\exp\left\{-(2\sigma^2)^{-1}(x-\mu)^2\right\}\\ &= \exp\left\{\kappa_1(\mu,\sigma^2)T_1(x) + \kappa_2(\mu,\sigma^2)T_2(x) - (1/2)[\mu^2/\sigma^2 + \ln(2\pi\sigma^2)]\right\},\end{aligned}$$

where

$$\begin{aligned}(T_1(x),T_2(x)) &= \left(x,x^2\right),\\ (\theta_1,\theta_2) &= \left(\kappa_1(\mu,\sigma^2),\kappa_2(\mu,\sigma^2)\right) := \left(\mu/\sigma^2, -1/(2\sigma^2)\right),\\ \left(\mu,\sigma^2\right) &= (\gamma_1(\theta),\gamma_2(\theta)) := (-\theta_1/(2\theta_2), -1/(2\theta_2)).\end{aligned} \tag{1.16}$$

Hence $\mathsf{N}(\mu,\sigma^2)$, $(\mu,\sigma^2)\in\mathbb{R}\times(0,\infty)$, is a reparametrized exponential family with generating statistic $T = (T_1,T_2)$ and φ_{μ,σ^2} turns into

$$\begin{aligned}f_\theta(x) &= \exp\left\{\theta_1 T_1(x) + \theta_2 T_2(x) - K(\theta)\right\}, \quad \text{where}\\ K(\theta) &= -(1/2)[-\theta_1^2/(2\theta_2) + \ln(-\theta_2/\pi)], \quad \theta\in\mathbb{R}\times(-\infty,0).\end{aligned} \tag{1.17}$$

The set $\Delta = \mathbb{R}\times(-\infty,0)$ is the natural parameter set as

$$\int \exp\left\{\theta_1 T_1(x) + \theta_2 T_2(x)\right\}\boldsymbol{\lambda}(dx) < \infty$$

if and only if $(\theta_1,\theta_2)\in\mathbb{R}\times(-\infty,0)$. Thus we have a regular two-parameter exponential family, represented in natural form by f_θ, $\theta\in\mathbb{R}\times(-\infty,0)$, and represented in reparametrized form by φ_{μ,σ^2}, $(\mu,\sigma^2)\in\mathbb{R}\times(0,\infty)$. The latter is based on the statistically relevant parameters μ and σ^2.

Suppose now that we have a sample of size n; i.e., let $X_1,...,X_n$ be i.i.d. with distribution $\mathsf{N}(\mu,\sigma^2)$. Then by Proposition 1.4 $\mathsf{N}^{\otimes n}(\mu,\sigma^2)$, $(\mu,\sigma^2)\in\mathbb{R}\times(0,\infty)$, is again an exponential family, but now with the generating statistic

$$T_{\oplus n}(x_1,...,x_n) = (\sum\nolimits_{i=1}^n x_i, \sum\nolimits_{i=1}^n x_i^2).$$

Problem 1.12.* The family of distributions in the reduced model $(\mathsf{N}^{\otimes n}(\mu,\sigma^2))\circ T_{\oplus n}^{-1}$ has the Lebesgue density $\sigma^{-2}\varphi_{n\mu,n\sigma^2}(s_1)\mathsf{h}_{n-1}(s_2/\sigma^2 - s_1^2/(n\sigma^2))$, where h_{n-1} is the Lebesgue density of a χ^2-distribution with $n-1$ degrees of freedom.

Next we consider some exponential families that appear as distributions of nonnegative random variables.

Example 1.13. Let $(\mathsf{Ga}(\lambda,\beta))_{\lambda,\beta>0}$ be the family of gamma distributions which have the Lebesgue densities

$$\mathsf{ga}_{\lambda,\beta}(x) = \frac{\beta^\lambda}{\Gamma(\lambda)} x^{\lambda-1}\exp\{-\beta x\} I_{(0,\infty)}(x), \quad x\in\mathbb{R}, \quad \lambda,\beta>0.$$

We introduce the measure $\boldsymbol{\mu}$ by $\boldsymbol{\mu}(dx) = I_{(0,\infty)}(x)\, x^{-1}\boldsymbol{\lambda}(dx)$, and set $T_1(x)=\ln x$, $T_2(x)=-x$, $x>0$. The $\boldsymbol{\mu}$-density is then given by (1.6), with $K(\lambda,\beta)=\ln\Gamma(\lambda)-\lambda\ln\beta$, and $(\mathsf{Ga}(\lambda,\beta))_{\lambda,\beta>0}$ becomes a two-parameter exponential family in natural form with natural parameter $\theta=(\lambda,\beta)\in\Delta=(0,\infty)\times(0,\infty)$ and generating statistic $T(x)=(\ln x,-x)$.

Problem 1.14. Represent the family $(\mathsf{Ga}(\lambda,\beta))_{\lambda,\beta>0}$ for a fixed known λ, as well as for a fixed known β, as a one-parameter exponential family in natural form. Extend this representation to the case of an i.i.d. sample $X_1,...,X_n$ where the distribution of X_1 belongs to the gamma family.

Problem 1.15.* Let $(\mathsf{Be}(\alpha,\beta))_{\alpha,\beta>0}$ be the family of beta distributions, which have the Lebesgue densities

$$\mathsf{be}_{\alpha,\beta}(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} I_{(0,1)}(x), \quad x\in\mathbb{R}, \quad \alpha,\beta>0.$$

It can be represented as a two-parameter exponential family in natural form.

Perhaps the most important and useful analytic property of an exponential family in natural form is that in its expectations, differentiations with respect to the coordinates of $\theta=(\theta_1,...,\theta_d)\in\Delta^0$ and integration with respect to $x\in\mathcal{X}$ can be exchanged. Denote by $\mathbb{C}=\{z: z=u+iv,\ u,v\in\mathbb{R}\}$ the set of complex numbers, where u and v are the real and imaginary parts of z, respectively. Similarly, we set $\mathbb{C}^d=\{z: z=u+iv,\ u,v\in\mathbb{R}^d\}$ and again denote by u and v the real and imaginary parts of the vector $z\in\mathbb{C}^d$. A function $\psi=\psi_1+i\psi_2$ is called measurable if ψ_1 and ψ_2 are real-valued measurable functions and denote this again by $\psi:\mathcal{X}\to_m\mathbb{C}$. We set for any $\boldsymbol{\mu}\in\mathcal{M}(\mathfrak{A})$

$$\mathbb{U}=\{u: u\in\mathbb{R}^d, \quad \int|\psi(x)|\exp\{\langle u,T(x)\rangle\}\boldsymbol{\mu}(dx)<\infty\}.$$

Then with $z=u+iv$ the relation $|\exp\{i\alpha\}|=1$ yields $|\exp\{\langle z,T(x)\rangle\}| = |\exp\{\langle u,T(x)\rangle\}|$ so that the function

$$M_\psi(z)=\int\psi(x)\exp\{\langle z,T(x)\rangle\}\boldsymbol{\mu}(dx)$$

is well defined on $\mathbb{F} = \mathbb{U} + i\mathbb{R}^d = \{z : z = u + iv,\, u \in \mathbb{U},\, v \in \mathbb{R}^d\}$. For brevity, we introduce the notation

$$D^\alpha \;:=\; \frac{\partial^{m_1+\cdots+m_d}}{\partial z_1^{m_1}\cdots\partial z_d^{m_d}}, \quad \alpha = (m_1, ..., m_d) \in \mathbb{N}^d,$$
$$|\alpha| = \sum\nolimits_{l=1}^{d} m_l, \quad z^\alpha = z_1^{m_1}\cdots z_d^{m_d}.$$

We recall that for an open set $A \subseteq \mathbb{C}^d$ a function $f : A \to \mathbb{C}$, is called *analytic* if, for every $z_0 \in A$, f can be expanded in a power series

$$f(z) = \sum\nolimits_{k=0}^{\infty} \sum_{\alpha:|\alpha|=k} \frac{1}{m_1!\cdots m_d!} a_\alpha (z - z_0)^\alpha$$

which is absolutely convergent in some neighborhood of z_0. In this case f is infinitely often differentiable and it holds

$$D^\alpha f(z_0) = a_\alpha. \tag{1.18}$$

The following result has been established in the literature in several different versions. Presumably, the first proof was presented in Lehmann (1959).

Lemma 1.16. *For every $\theta_0 \in \mathbb{U}^0$ there exists some $\varepsilon > 0$ such that*

$$\int \exp\{\varepsilon \,\|T(x)\|\}|\psi(x)|\exp\{\langle\theta_0, T(x)\rangle\}\boldsymbol{\mu}(dx) < \infty. \tag{1.19}$$

The function $M_\psi(z) = \int \psi(x)\exp\{\langle z, T(x)\rangle\}\boldsymbol{\mu}(dx)$ is analytic in the interior $\mathbb{F}^0 = \mathbb{U}^0 + i\mathbb{R}^d$ of $\mathbb{F}$, and it holds for $\alpha = (m_1, ..., m_d)$,

$$\begin{aligned} D^\alpha M_\psi(z) &= \int \psi(x) T_1^{m_1}(x)\cdots T_d^{m_d}(x)\exp\{\langle z, T(x)\rangle\}\boldsymbol{\mu}(dx) \\ &= \int \psi(x) D^\alpha \exp\{\langle z, T(x)\rangle\}\boldsymbol{\mu}(dx),\ z \in \mathbb{F}^0. \end{aligned}$$

Proof. Fix $z_0 = (z_{1,0}, ..., z_{d,0}) \in \mathbb{U}^0 + i\mathbb{R}^d$ and $z = (z_1, ..., z_d)$ and denote by u_i and $u_{i,0}$ the real parts of z and z_0, respectively. The inequalities $\|T\| \le \sum_{i=1}^d |T_i|$ and $\exp\{|x|\} \le \exp\{x\} + \exp\{-x\}$ imply (1.19). The latter inequality and $\sum_{k=0}^n |w|^k/k! \le \exp\{|w|\}$ yields for $\|z - z_0\| \le \delta$

$$\begin{aligned} &|\psi(x)\exp\{\langle z_0, T(x)\rangle\}|\sum\nolimits_{l=0}^{n} \frac{|\langle z - z_0, T(x)\rangle|^l}{l!} \\ &\le |\,\psi(x)\exp\{\sum\nolimits_{j=1}^{d} u_{j,0}T_j(x)\}\,|\,\exp\{\delta\sum\nolimits_{j=1}^{d}|T_j(x)|\} \\ &\le \sum_{\varepsilon_1,...,\varepsilon_d\in\{-1,0,1\}} |\psi(x)|\exp\{\sum\nolimits_{j=1}^{d}(u_{j,0} + \varepsilon_j\delta)T_j(x)\}. \end{aligned}$$

For sufficiently small δ and the vectors $(u_{1,0}+\varepsilon_1\delta, ..., u_{d,0}+\varepsilon_d\delta)$ belong to $\mathbb{U}^0$ so that the function on the right-hand side of the above inequality is integrable with respect to $\boldsymbol{\mu}$. Hence by Lebesgue's theorem (see Theorem A.18),

$$\begin{aligned} M_\psi(z) &= \int \psi(x) \exp\{\langle z_0, T(x)\rangle\} \sum\nolimits_{k=0}^{\infty} \frac{\langle z - z_0, T(x)\rangle^k}{k!} \boldsymbol{\mu}(dx) \\ &= \sum\nolimits_{k=0}^{\infty} \sum_{|\alpha|=k} \frac{1}{m_1! \cdots m_d!} a_\alpha (z - z_0)^\alpha, \end{aligned}$$

where

$$a_\alpha = \int \psi(x) T_1^{m_1}(x) \cdots T_1^{m_d}(x) \exp\{\langle z_0, T(x)\rangle\} \boldsymbol{\mu}(dx).$$

The relation $a_\alpha = D^\alpha f(z_0)$ in (1.18) with $f = M_\psi$ completes the proof. ■

Theorem 1.17. *Let $(P_\theta)_{\theta\in\Delta}$ be an exponential family in natural form as given by* (1.5)*. Then for every $\theta \in \Delta^0$ there exists an $\varepsilon > 0$ with*

$$\mathsf{E}_\theta \exp\{\varepsilon \|T\|\} < \infty, \tag{1.20}$$

so that

$$\mathsf{E}_\theta \|T\|^a < \infty \quad \textit{for every } a > 0. \tag{1.21}$$

The function K is infinitely often differentiable in Δ^0 and it holds for every $\alpha = (m_1, ..., m_d) \in \mathbb{N}^d$,

$$\mathsf{E}_\theta T_1^{m_1} \cdots T_d^{m_d} = \exp\{-K(\theta)\} D^\alpha \exp\{K(\theta)\}. \tag{1.22}$$

Proof. The statement (1.20) follows from (1.19), and (1.21) is implied by (1.20). From Lemma 1.16 we can see for $\psi(x) \equiv 1$ that the real-valued function $\exp\{K(\theta)\} = \int \exp\{\langle\theta, T\rangle\} d\boldsymbol{\mu}$ is infinitely often differentiable in $\mathbb{U}^0 = \Delta^0$. Because $\int \exp\{\langle\theta, T\rangle\} d\boldsymbol{\mu} \neq 0$, the function $K(\theta)$ is also infinitely often differentiable. To prove (1.22) we note that by Lemma 1.16

$$\begin{aligned} &\exp\{K(\theta)\} \mathsf{E}_\theta T_1^{m_1} \cdots T_d^{m_d} = \int T_1^{m_1}(x) \cdots T_d^{m_d}(x) \exp\{\langle\theta, T(x)\rangle\} \boldsymbol{\mu}(dx) \\ &= \int D^\alpha \exp\{\langle\theta, T(x)\rangle\} \boldsymbol{\mu}(dx) = D^\alpha \int \exp\{\langle\theta, T(x)\rangle\} \boldsymbol{\mu}(dx) \\ &= D^\alpha \exp\{K(\theta)\}. \end{aligned}$$

■

Remark 1.18. In the previous lemma we have proved the existence of all moments of T provided the parameter belongs to the interior of Δ. For the boundary points, in general, this statements is no longer true as the following example shows. Consider the inverse Gaussian distribution $\mathsf{Gi}(\lambda, m)$ with natural parameters $(\theta_1, \theta_2) = \big(-\lambda/(2m^2), -\lambda/2\big) \in (-\infty, 0] \times (-\infty, 0)$ and Lebesgue density $\mathsf{gi}_{\lambda,m}$ from (1.14) for $\theta_1 = 0$, i.e., $\mathsf{gi}_{\lambda,\infty}$ in (1.15). Obviously $\mathsf{E}_{0,\theta_2} T_1 = \int_0^\infty x \; \mathsf{gi}_{\lambda,\infty}(x) dx = \infty$.

There is a simple explanation for this effect. We have pointed out in Example 1.10 that $\mathsf{gi}_{\lambda,m}$ is the density of the first passage time at which the process $\nu t + \sigma W(t)$ crosses the level a, where $\lambda = (a/\sigma)^2$ and $m = a/\nu$. The case $m = \infty$ corresponds to $\nu = 0$, i.e., there is no positive drift. In this case the Wiener process hits the level a very late so that the hitting time is finite with probability one, but the expected value is infinite.

For brevity, we introduce the notation

$$\nabla = \left(\frac{\partial}{\partial\theta_1}, ..., \frac{\partial}{\partial\theta_d}\right)^T \quad \text{and} \quad \nabla\nabla^T = \left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\right)_{1\le i,j\le d}.$$

The following formulas for calculating the means and covariances of $T_1, ..., T_d$ are direct consequences of (1.22).

Corollary 1.19. *Under the assumptions of Theorem* 1.17*, and conditions (A1) and (A2), for every $\theta \in \Delta^0$ the mean vector and the covariance matrix of T are given by*

$$\mathsf{E}_\theta T = \nabla K(\theta), \quad \mathsf{C}_\theta(T) = \nabla\nabla^T K(\theta). \tag{1.23}$$

The matrix $\nabla\nabla^T K(\theta)$ is nonsingular for every $\theta \in \Delta^0$ and the infinitely often differentiable function K is strictly convex in Δ^0.

Proof. Let $\theta \in \Delta^0$. From (1.22) we get for any $\theta \in \Delta^0$ and $D^\alpha = \frac{\partial}{\partial\theta_i}$,

$$\mathsf{E}_\theta T_i = \exp\{-K(\theta)\}\frac{\partial}{\partial\theta_i}\exp\{K(\theta)\} = \frac{\partial K(\theta)}{\partial\theta_i}.$$

This proves the first statement. Similarly with $D^\alpha = \frac{\partial^2}{\partial\theta_i\partial\theta_j}$,

$$\begin{aligned}\mathsf{E}_\theta T_i T_j &= \exp\{-K(\theta)\}\frac{\partial^2}{\partial\theta_i\partial\theta_j}\exp\{K(\theta)\} \\ &= \frac{\partial K(\theta)}{\partial\theta_i}\frac{\partial K(\theta)}{\partial\theta_j} + \frac{\partial^2 K(\theta)}{\partial\theta_i\partial\theta_j} = \frac{\partial K(\theta)}{\partial\theta_i}\frac{\partial K(\theta)}{\partial\theta_j} + \frac{\partial^2 K(\theta)}{\partial\theta_i\partial\theta_j}.\end{aligned}$$

The nonsingularity of $\nabla\nabla^T K(\theta)$ follows from $\mathsf{C}_\theta(T) = \nabla\nabla^T K(\theta)$ and the fact that by assumption (A1) the components of T are not a.s. linearly dependent, and Problem 1.3. We already know from (1.4) that K is convex. The nonsingularity of $\nabla\nabla^T K(\theta)$ implies that K is strictly convex. ■

We illustrate the above results by examples.

Example 1.20. It has been shown in Example 1.13 that $(\mathsf{Ga}(\alpha,\beta))_{\alpha,\beta>0}$ is a two parameter exponential family in natural form with natural parameter (λ,β) and generating statistic $T(x) = (T_1(x), T_2(x))$, where by $K(\lambda,\beta) = \ln\Gamma(\lambda) - \lambda\ln\beta, \lambda,\ \beta > 0$. From (1.23) we get, with $\Psi = \Gamma'/\Gamma$,

$$\mathsf{E}_\theta T = (\Psi(\lambda) - \ln\beta, -\frac{\lambda}{\beta}) \quad \text{and} \quad \mathsf{C}_\theta(T) = \begin{pmatrix} \Psi'(\lambda) & -\frac{1}{\beta} \\ -\frac{1}{\beta} & \frac{\lambda}{\beta^2} \end{pmatrix}.$$

Example 1.21. Let $X_1, X_2, ...$ be a Bernoulli sequence with success probability p, where $p \in (0,1)$. For a fixed given $k \in \{1,2,...\}$ let $X = \min\{n : X_1 + \cdots + X_n = k\} - k$. Thus, $X + k$ is the number of times one has to play a game with winning probability p in independent repetitions until k games have been won. X follows a negative binomial distribution $\mathsf{Nb}(k,\eta)$ with p.m.f.

$$\mathsf{nb}_{k,p}(x) = \frac{(x+k-1)!}{x!(k-1)!} p^k (1-p)^x, \quad x = 0,1,2,...$$

Put $\theta = \kappa(p) = \ln(1-p)$, and $\boldsymbol{\mu}(\{x\}) = (x+k-1)!/(x!(k-1)!)$. Then the distribution of X has the density $f_\theta(x) = \exp\{\theta x + k\ln(1-\exp\{\theta\})\}$ with respect to $\boldsymbol{\mu}$. This shows that $\mathsf{Nb}(k, 1-e^\theta)$ is a one-parameter exponential family with $T(x) = x$ and $K(\theta) = -k\ln(1-\exp\{\theta\})$. From (1.23) we get

$$\mathsf{E}_{\kappa(p)} T = k\frac{1-p}{p} \quad \text{and} \quad \mathsf{V}_{\kappa(p)}(T) = k\frac{1-p}{p^2}.$$

In the previous examples we have already studied different ways of parametrizing an exponential family. However, among all parametrizations there is one in particular, not mentioned so far, that has a special meaning. This is the so-called *mean value parametrization* which is considered at the conclusion of this section. To prepare for this parametrization we need the following well-known result (see, e.g., Brown (1986) and Witting (1985)).

Theorem 1.22. *Under the assumptions of (A1) and (A2) the mapping*

$$\gamma_{\mathsf{m}} : \theta \mapsto \nabla K(\theta) = \mathsf{E}_\theta T \tag{1.24}$$

is a diffeomorphism of Δ^0 onto the open set $\gamma_{\mathsf{m}}(\Delta^0)$.

Proof. We already know from Corollary 1.19 that K is strictly convex. This yields for every $\theta_1, \theta_2 \in \Delta^0$ with $\theta_1 \neq \theta_2$,

$$\begin{aligned} K(\theta_1) &> K(\theta_2) + \langle(\theta_1 - \theta_2), \nabla K(\theta_2)\rangle\,, \\ K(\theta_2) &> K(\theta_1) + \langle(\theta_2 - \theta_1), \nabla K(\theta_1)\rangle\,. \end{aligned}$$

Hence, $\langle \theta_2 - \theta_1, \nabla K(\theta_1) - \nabla K(\theta_2)\rangle < 0$, so that $\theta_1 \neq \theta_2$ implies $\nabla K(\theta_1) \neq \nabla K(\theta_2)$ and γ_{m} is a bijection. As by Proposition 1.16 K is infinitely often differentiable we see that the mapping γ_{m} is continuously differentiable. An application of the global inverse function theorem (see, e.g., Theorem 3.2.8 in Duistermaat and Kolk (2004)) completes the proof. ■

Brown (1986) proved under the so-called steepness condition a stronger result which at the same time characterizes the range $\gamma_{\mathsf{m}}(\Delta^0)$. We come back to this result later when we study maximum likelihood estimators in exponential families in Section 7.5.

By denoting the inverse mapping of γ_{m} by κ_{m}, we can represent the exponential family in the mean value parametrization, at least for $\theta \in \Delta^0$, by

$$P_{\kappa_{\mathsf{m}}(\mu)}, \quad \mu \in \gamma_{\mathsf{m}}(\Delta^0).$$