WILEY SERIES IN PROBABILITY AND STATISTICS

Information and Exponential Families in Statistical Theory

Ole E. Barndorff-Nielsen

WILEY

Information and Exponential Families

Information and Exponential Families

In Statistical Theory

O. BARNDORFF-NIELSEN
Matematisk Institut
Aarhus Universitet

John Wiley & Sons Chichester · New York · Brisbane · Toronto

This edition first published 2014

Registered office
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

Library of Congress Cataloging-in-Publication Data

Barndorff-Neilsen, Ole
Information and exponential families

(Wiley series in probability and mathematical statistics –
tracts)

Includes bibliographical references and index.

1. Sufficient statistics. 2. Distribution (Probability theory).
3. Functions, Exponential. I. Title.
QA276.B2847 519.5 77-9943
ISBN 0 471 99545 2

A catalogue record for this book is available from the British Library.

ISBN: 978-1-118-85750-2

2 2014

A Note from the Author

This book - a reprint of the original from 1978 - provides a systematic discussion of the basic principles of statistical inference. It was written at the time near the end of the forty years period where, starting with R. A. Fisher's seminal work, the debate about such principles was very active. Since then there have been no major developments, reflecting the fact that the core principles are accepted as valid and seem effectively exhaustive. These core principles are likelihood, sufficiency and ancillarity, and various aspects of these.

The theory is illustrated with numerous examples, of both theoretical and applied interest, some of them arising from concrete questions in other fields of science.

Extensive comments and references to the relevant literature are given in notes at the end of the separate chapters.

The account of the principles of statistical inference constitutes Part I of the book. Part II presents some technical material, needed in Part III for the exposition of the exact, as opposed to asymptotic, theory of exponential families, as it was known at the time. Some discussion is also given of transformation families, a subject that was still in its early stages then. The exact properties are related to the inference principles discussed in Part I.

Much of the later related work has been concerned with approximate versions of the core principles and of associated results, for instance about the distribution of the maximum likelihood estimator. A few references to related subsequent work are given below.

Exponential transformation models. (1982) Proc. Roy. Soc. London A 379, 41-65; coauthors Blæsild, P., Jemsen, J.L. and Jørgensen, B..; On a formula for the distribution of the maximum likelihood estimator. (1983) Biometrika 73, 307-322.; Likelihood Theory. Chapter 10 in D.V. Hinkley, N. Reid and Snell, E.J. Statistical Theory and Modelling. (1983) London; Chapman and Hall.; Inference on full and partial parameters, based on the standardized log likelihood ratio. (1983) Biometrika 73, 307-322.; Parametric Statistical Models and Likelihood. (1988). Springer Lecture Notes in Statistics. Heidelberg: Springer-Verlag; Inference and Asymptotics .London: Chapman and Hall.(1994); coauthor Cox, D.R; General exponential families. Encyclopedia of Statistical Sciences. (1997) Update Volume 1, 256-261.

October 2013
Ole E. Barndorff-Nielsen

Preface

This treatise brings together results on aspects of statistical information, notably concerning likelihood functions, plausibility functions, ancillarity, and sufficiency, and on exponential families of probability distributions. A brief outline of the contents and structure of the book is given in the beginning of the introductory chapter.

Much of the material presented is of fairly recent origin, and some of it is new. The book constitutes a further development of my Sc.D. thesis from the University of Copenhagen (Barndorff-Nielsen 1973a) and includes results from a number of my later papers as well as from papers by many other authors. References to the literature are given partly in the text proper, partly in the Notes sections at the ends of Chapters 2–4 and 8–10.

The roots of the book lie in the writings of R. A. Fisher both as concerns results and the general stance to statistical inference, and this stance has been a determining factor in the selection of topics.

Figures 2.1 and 10.1 are reproduced from Barndorff-Nielsen (1976a and b) by permission of the Royal Statistical Society, and Figures 10.2 and 10.3 are reproduced from Barndorff-Nielsen (1973c) by permission of the Biometrika Trustees. The results from R. T. Rockafellar's book *Convex Analysis* (copyright © 1970 by Princeton University Press) quoted in Chapter 5 are reproduced by permission of Princeton University Press.

In the work I have benefited greatly from discussions with colleagues and students. Adding to the acknowledgements in my Sc.D. thesis, I wish here particularly to express my warm gratitude to Preben Blæsild, David R. Cox, Jørgen G. Pedersen, Helge Gydesen, Geert Schou, and especially Anders H. Andersen for critical readings of the manuscript, to David G. Kendall for helpful and stimulating comments, and to Anne Reinert for unfailingly excellent and patient secretarial assistance. A substantial part of the manuscript was prepared in the period August 1974–January 1975 which I spent in Cambridge, at Churchill College and the Statistical Laboratory of the University. I am most grateful to my colleagues at the Department of Theoretical Statistics, Aarhus University, and the Statistical Laboratory, Cambridge University, and to the Fellows of Churchill for making this stay possible.

Aarhus, May 1977 O. B. -N.

Contents

Part II
Convex analysis, unimodality, and Laplace transforms

Part III
Exponential families

CHAPTER 1

Introduction

1.1 INTRODUCTORY REMARKS AND OUTLINE

The main kinds of task in statistics are the construction or choice of a statistical model for a given set of data, and the assessment and charting of statistical information in model and data.

This book is concerned with certain questions of statistical information thought to be of interest for purposes of scientific inference. It also contains an account of the theory of exponential families of probability measures, with particular reference to those questions. Besides exponential families, the most important type of statistical models are the group families, i.e. families of probability measures generated by a unitary group of transformations on the sample space. However, only the most basic facts on group families will be referred to. (Some further introductory remarks on these two types of models are given in Section 1.3.) Another limitation is that asymptotic problems are not discussed, except for a few remarks.

The reader is supposed to have a fairly broad, basic knowledge of statistical inference, and in particular to be familiar with the more conceptual aspects of likelihood and plausibility, such as are discussed in Birnbaum (1969) and Barndorff-Nielsen (1976b), respectively.

Probability functions, likelihood functions, and plausibility functions are charts of different types of statistical information. They are the three prominent instances of the concept of ods functions, due to Barnard (1949). An ods function is a real function on the space of possible experimental outcomes or on the space of hypotheses, which expresses the relative 'credibility' of the points of the space in question. It is often convenient to work with the logarithms of such functions and these are termed lods functions. For the objectives of this treatise the interest in lods (or ods) functions lies mainly in the very concept which is instrumental in bringing to the fore the duality between the sample aspect and the parameter aspect of statistical models, and in constructing prediction functions. Thus, although the concept of lods function will be referred to at a number of places, the theoretical developments relating to lods functions and presented in Barnard (1949) are not of direct relevance in the present context and will only be indicated briefly (in Section 3.1).

Generally, only part of the statistical information contained in the model and the data is pertinent to a given question, and one is then faced with the problem of separating out that part. The key procedures for such separations are margining to a sufficient statistic and conditioning on an ancillary statistic. Basic here is the concept of nonformation, i.e. the concept that a certain submodel and the corresponding part of the data contain no (accessible) pertinent or relevant information in respect of the question of interest.

A general treatment of the topics of statistical information indicated above is given in Part I, while the theory of exponential families is developed in Part III. Properties of convex sets and functions, in particular convex duality relations, are of great importance for the study of exponential families. Since much of convex analysis is of fairly recent origin and is not common knowledge, a compendious account of the relevant results is given in Part II, together with properties of unimodality and Laplace transforms. A reader primarily interested in lods functions and exponential families may concentrate on Chapters 2, 3, 8, and 9, just referring to Part II, which consists of Chapters 5–7, as need arises. Inferential separation, hereunder notably nonformation, ancillarity, and sufficiency, is discussed in Chapters 4 and 10. The chapters of Parts I and III contain Complements sections where miscellaneous results which did not fit into the mainstream of the text have been collected.

Each known methodological approach, of any inclusiveness, to the questions of statistical inference is hampered by various difficulties of logical or epistemic character, and applications of these approaches must therefore be tempered by independent judgement. The merits of any one approach depend on the extent to which it yields sensible and useful answers as well as on the cogency of its fundamental ideas.

The difficulties, of the kind mentioned, connected with likelihood, plausibility, ancillarity, and sufficiency have been discussed in Birnbaum (1969), Barndorff–Nielsen (1976b), and numerous other papers. Many of these papers will be referred to in the course of this treatise, but a comprehensive exposition of the arguments adduced will not be given. One of the difficulties, whose seriousness seems to have been overestimated, is that different applications of ancillarity and sufficiency, to the same model and data, may lead to different inferential conclusions (cf. Section 4.7(**vi**)). However, as has been stressed and well illustrated by Barnard (1974b), it is in general impossible to obtain unequivocal conclusions on the basis of statistical information. It is therefore not surprising that if uniqueness in conclusions is presupposed as a requirement of inference then paradoxical results turn up, such as is the case with Birnbaum's Theorem (Section 4.7(v)).

1.2 SOME MATHEMATICAL PREREQUISITES

Let M be a subset of a space $\mathfrak{M}$. The *indicator* of M is the function 1_M defined by

$$1_M(x) = \begin{cases} 1 & \text{for } x \in M \\ 0 & \text{for } x \in M^c \end{cases}$$

where M^c is the complement $\mathfrak{M} \backslash M$ of M. If $\mathfrak{M}$ is a product space, $\mathfrak{M} = \mathfrak{M}_1 \times \mathfrak{M}_2$, and if $x_1 \in \mathfrak{M}_1$ then M_{x_1} is the section of M at x_1, i.e. $M_{x_1} = \{x_2 : (x_1, x_2) \in M\}$. When $\mathfrak{M}$ is a topological space the interior, closure, and boundary of M are denoted by int M, cl M, and bd M, respectively. Suppose $\mathfrak{M} = R^k$. The affine hull of M is written aff M, and dim M is the dimension of aff M. An affine subset of M is a set of the form $M \cap L$ where L is an affine subspace of R^k.

For any mapping f the notations domain f and range f will be used, respectively, for the domain of definition of f and the range of f, and f is said to be a mapping *on* domain f.

If x is a real number then $[x\}$ will stand for $x - 1$ or x provided x is an integer and for $[x]$, the integer part of x, otherwise. Furthermore, the notations $N = \{1, 2, \ldots\}$, $N_0 = \{0, 1, 2, \ldots\}$, and $Z = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$ are adopted.

All vectors are considered basically as row vectors, and the length of a vector x is indicated by $|x|$. A set of vectors in R^k are said to be *affinely independent* provided their endpoints do not belong to an affine subspace of R^k. The transpose of a matrix $\mathbf{A}$ is denoted by $\mathbf{A}'$ and, for $\mathbf{A}$ quadratic, $|\mathbf{A}|$ is the determinant and tr $\mathbf{A}$ is the trace of $\mathbf{A}$. The symbols $\mathbf{I}$ or $\mathbf{I}_r$ are used for the $r \times r$ unit matrix ($r = 1, 2, \ldots$). Occasionally an $r \times r$ symmetric matrix $\mathbf{A}$ with elements a_{ij}, say, will be interpreted either as a point in R^{r^2} or as the point in $R^{\binom{r+1}{2}}$ whose coordinates are given by $(a_{11}, a_{22}, \ldots, a_{rr}, a_{12}, \ldots, a_{1r}, a_{23}, \ldots, a_{2r}, \ldots, a_{r-1,r})$. Let $\mathbf{\Sigma}$ be a positive definite matrix, set $\mathbf{\Delta} = \mathbf{\Sigma}^{-1}$, and let

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix} \quad \text{and} \quad \mathbf{\Delta} = \begin{pmatrix} \mathbf{\Delta}_{11} & \mathbf{\Delta}_{12} \\ \mathbf{\Delta}_{21} & \mathbf{\Delta}_{22} \end{pmatrix}$$

be similar partitions of $\mathbf{\Sigma}$ and $\mathbf{\Delta}$. Then, as is well known,

$$\mathbf{\Delta}_{22}^{-1} = \mathbf{\Sigma}_{22} - \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12} \tag{1}$$

$$-\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12} = \mathbf{\Delta}_{12}\mathbf{\Delta}_{22}^{-1}. \tag{2}$$

When indexed variables, as for example $x_i, i = 1, \ldots, m$, or $x_{ij}, i = 1, \ldots, m$; $j = 1, \ldots, n$, are considered the substitution of a dot for an index variable signifies summation over that variable. Furthermore, the vector $(x_1, \ldots, x_m)$ will be denoted by x_*, the vector $(x_{i1}, \ldots, x_{in})$ by x_{i*}, etc.

Consider a real-valued function f defined on a subset $\mathfrak{X}$ of R^k. The notations $Df = \partial f / \partial x$ and $\partial^2 f / \partial x' \partial x$ are used for the gradient and the matrix of second order derivatives of f, respectively, while $D^i f$, where $i = (i_1, \ldots, i_k)$ is a vector of non-negative integers, indicates a mixed derivative of f. (Thus $Df = D^{(1,\ldots,1)}f$.) In the case where a partition $(x^{(1)}, \ldots, x^{(m)})$ of $x (\in \mathfrak{X})$ is given then the (i, j)th matrix component of the corresponding partition of $\partial^2 f / \partial x' \partial x$ is denoted by $\partial^2 f / \partial x^{(i)\prime} \partial x^{(j)}$. Let h be a twice continuously differentiable mapping on an open

subset $\mathfrak{Y}$ of R^k and onto $\mathfrak{X}$, also assumed open, and set

$$\frac{\partial x}{\partial y'} = \frac{\partial h}{\partial y'} = \begin{pmatrix} \frac{\partial h_1}{\partial y_1} & \cdots & \frac{\partial h_k}{\partial y_1} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \frac{\partial h_1}{\partial y_k} & \cdots & \frac{\partial h_k}{\partial y_k} \end{pmatrix}$$

the Jacobian matrix of h. Moreover, set

$$\frac{\partial^2 x}{\partial y' \partial y} = \frac{\partial^2 h}{\partial y' \partial y} = \left(\frac{\partial^2 h_1}{\partial y' \partial y}, \ldots, \frac{\partial^2 h_k}{\partial y' \partial y} \right).$$

If f is twice continuously differentiable then, writing $\tilde{f}$ for the composition $f \circ h$ of f and h, one has

$$\frac{\partial^2 \tilde{f}}{\partial y' \partial y} = \frac{\partial x}{\partial y'} \frac{\partial^2 f}{\partial x' \partial x} \frac{\partial x'}{\partial y} + \frac{\partial f}{\partial x} \cdot \frac{\partial^2 x}{\partial y' \partial y} \tag{3}$$

where $\cdot$ is a matrix multiplication symbol defined in the following way. For a $1 \times k$ vector $v = (v_1, \ldots, v_k)$ and an $m \times nk$ matrix $\mathbf{A} = [\mathbf{A}_1, \ldots, \mathbf{A}_k]$, $\mathbf{A}_i$ being $m \times n$ $(i = 1, \ldots, k)$, the product $v \cdot \mathbf{A}$ is given by

$$v \cdot \mathbf{A} = v_1 \mathbf{A}_1 + \cdots + v_k \mathbf{A}_k.$$

(Thus the operation $\cdot$ is a generalization of the ordinary inner product of two k-dimensional vectors.)

Measure-theoretic questions concerning null sets, measurability of mappings, etc., will largely be bypassed. (Section 4.2, however, forms something of an exception to this.) The mathematical gaps left thereby may be filled out by standard reasoning.

Lebesgue measure will be denoted by λ, counting measure by ν. (The domains of these measures vary from case to case but it will be apparent from the context what the domain is.)

Let H be a class of transformations on a space $\mathfrak{X}$, i.e. the elements of H are one-to-one mappings of $\mathfrak{X}$ onto itself. The class H is *unitary*, respectively *transitive*, if for every pair of points x and $\tilde{x}$ in $\mathfrak{X}$ the equation $\tilde{x} = h(x)$ has at most, respectively at least, one solution h in H. In the case where H is transitive, the set $H(x) = \{h(x): h \in H\}$ is equal to $\mathfrak{X}$. A measure μ on a σ-algebra $\mathfrak{A}$ of $\mathfrak{X}$ is transformation invariant under H if $\mu h = \mu$ for every $h \in H$, where μh is defined by $\mu h(A) = \mu(h^{-1}(A))$, $A \in \mathfrak{A}$. Suppose H is a group under the operation $\circ$ of composition of mappings. Then $H(x)$ is called the *orbit* of x and the orbits form a partition of $\mathfrak{X}$. If, in addition, H is unitary then each orbit can be brought into one-to-one correspondence with H, and thus $\mathfrak{X}$ can be represented as a product space

$\mathfrak{U} \times \mathfrak{B}$ of points (u, v) such that u determines the orbit and v the position on that orbit of the point x in $\mathfrak{X}$ corresponding to (u, v). As is well known (see e.g. Nachbin 1965), if H is a locally compact, topological group then there exist left invariant as well as right invariant measures on H. For H unitary and transitive, these measures can, by the above identification of $\mathfrak{X}$ and H, also be viewed as transformation invariant measures on $\mathfrak{X}$.

The sample spaces to be considered are exclusively *Euclidean*, i.e. they are Borel subsets of Euclidean spaces, and the associated σ-algebras of events are the classes of Borel subsets of the sample spaces. Moreover, all random variables and statistics take values in Euclidean spaces. Generally, the letter $\mathfrak{X}$ will be used to denote the sample space, and x is a point in $\mathfrak{X}$.

Ordinarily, the same notation—a lower case italic letter—will be used for a random variable or statistic and for its value, the appropriate interpretation being determined by the context. In cases where clarity demands a distinction the mapping is denoted by the capital version of the letter.

Let $\mathfrak{X}$ ($\subset R^k$) be a sample space, $\mathfrak{A}$ the σ-algebra of Borel subsets of $\mathfrak{X}$, and $\mathfrak{P}$ a family of probability measures on $\mathfrak{X}$. The triplet $(\mathfrak{X}, \mathfrak{A}, \mathfrak{P})$ is termed a *statistical field*. Let P be a member of $\mathfrak{P}$ and let t (also T) be a statistic.

The marginal distribution of t under P has probability measure Pt given by $Pt(B) = P(t^{-1}(B))$ for Borel sets B. Further, $E_P t$ and $V_P t$ stand for the mean value (vector) and the variance (matrix) of t. For an event A with $P(A) > 0$ the conditional probability measure given A is denoted by $P(\cdot|A)$ or P^A. If $\mathfrak{B}$ is a sub-σ-algebra of $\mathfrak{A}$ then $P_{\mathfrak{B}}$ denotes the restriction of P to $\mathfrak{B}$ and $P^{\mathfrak{B}}$ is the Markov kernel of the conditional distribution given $\mathfrak{B}$ under P. The conditional mean value given $\mathfrak{B}$ under P of a random variable y is written $E_P^{\mathfrak{B}} y$. When $\mathfrak{B}$ is the σ-algebra generated by a statistic t the notations P_t, P^t or $P(\cdot|t)$, and $E_P^t y$ are normally used instead of $P_{\mathfrak{B}}$, $P^{\mathfrak{B}}$, and $E_P^{\mathfrak{B}} y$. For any measure μ on $\mathfrak{X}$, let $\mu^{(n)}$ indicate the measure on the product space $\mathfrak{X}^n$ which is the n-fold product of μ with itself, and let $\mu^{(*n)}$ be the n-fold convolution of μ (provided it exists). Set $\mathfrak{P}_t = \{P_t : P \in \mathfrak{P}\}$, $\mathfrak{P}^t = \{P^t : P \in \mathfrak{P}\}$, $\mathfrak{P}^{(n)} : P \in \mathfrak{P}\}$, etc.

If P and Q are two probability measures on $\mathfrak{X}$ having common support then

$$\frac{dP_t}{dQ_t} = E_Q^t \frac{dP}{dQ} \tag{4}$$

and

$$\frac{dP(\cdot|t)}{dQ(\cdot|t)} = \frac{dP/dQ}{dP_t/dQ_t}. \tag{5}$$

A distribution on R^k is *singular* if its affine support (i.e. the affine hull of its support) is a proper subset of R^k. Let u and v be statistics. The conditional distribution of u given v and under P is *singular* provided that the marginal distribution of u under the conditional probability measure given v is singular

with probability 1, i.e.

$$P\{Pu(\cdot|v) \text{ is singular}\} = 1.$$

The probability measure P is said to be of *discrete type* if the support S of P has no accumulation points, of *c-discrete type* if S equals the intersection of the set Z^k and some convex set, and of *continuous type* if P is absolutely continuous with respect to Lebesgue measure λ on $\mathfrak{X}$. The same terms are applied to $\mathfrak{P}$ provided each member of $\mathfrak{P}$ has the property in question.

A function ψ defined on $\mathfrak{P}$ and taking values in some Euclidean space is called a *parameter function.* As with random variables and statistics, the same notation ψ will generally be used for the function and its value, but when it seems necessary to distinguish explicitly the function is indicated by $\psi(\cdot)$. Suppose $\mathfrak{P}$ is given as an indexed set, $\mathfrak{P} = \{P_\omega : \omega \in \Omega\}$, then $\mathfrak{P}$ is called *parametrized* provided Ω is a subset of a Euclidean space and the mapping $\omega \to P_\omega$ is one-to-one. Any parameter function ψ on $\mathfrak{P}$ may be viewed as a function of ω, and its values will be denoted, freely, by $\psi(\omega)$ as well as by ψ or $\psi(P_\omega)$. Similarly for other kinds of mappings.

The family $\mathfrak{P}$ is said to be generated by a class H of transformations on $\mathfrak{X}$ if for some member P of $\mathfrak{P}$ one has $\mathfrak{P} = \{Ph : h \in H\}$. In the case where H is a unitary group the family $\mathfrak{P}$ will be called a *group family*. Suppose that $\mathfrak{P}$ is a group family and that u is a statistic which is constant on the orbits under H but takes different values on different orbits (thus u is a maximal invariant). Then u is said to *index* the orbits, and the marginal distribution of u is the same under all the elements of $\mathfrak{P}$. It is also to be noticed that if $\mathfrak{P}$ is a transitive group family (i.e. a group family with H transitive) and if μ is a left or right invariant measure on $\mathfrak{X}$ which, when interpreted as a transformation invariant measure on $\mathfrak{X}$, dominates $\mathfrak{P}$ then the family $\mathfrak{p}$ of probability functions or densities of $\mathfrak{P}$ relative to μ is of the form

$$\mathfrak{p} = \{p(h^{-1}(\cdot)) : h \in H\} \tag{6}$$

where $p = dP/d\mu$.

For the discussions in Parts I and III (except Section 3.1) it is presupposed that a statistical model, with sample space $\mathfrak{X}$ and family of probability measures $\mathfrak{P}$, has been formulated. Unless explicitly stated otherwise, it is moreover supposed that $\mathfrak{P}$ is parametrized, $\mathfrak{P} = \{P_\omega : \omega \in \Omega\}$, and determined by a family of probability functions $\mathfrak{p} = \{p(\cdot; \omega) : \omega \in \Omega\}$, i.e. $p(\cdot; \omega)$ is the density of P_ω with respect to a certain σ-finite measure μ which dominates $\mathfrak{P}$. For discrete-type distributions this dominating measure is always taken to be counting measure, so that $p(x; \omega)$ is the probability of x. (In subsequent chapters certain topics in plausibility inference will be considered. Whenever this is the case, it is—for non-discrete distributions—presupposed that $\sup_x p(x; \omega) < \infty$ for every $\omega \in \Omega$.) In the case $\mathfrak{p}$ is of the form (6), for some probability function p with respect to μ and some class H of transformations on $\mathfrak{X}$, then $\mathfrak{p}$ is said to be generated by H. The points x in $\mathfrak{X}$ for which $p(x; \omega) > 0$ for some $\omega \in \Omega$ are called *realizable*, and the *realizable values* of a statistic are the values corresponding to realizable sample points x.

Viewed as a function on $\mathfrak{X} \times \Omega$, $p(\cdot;\cdot)$ is referred to as the *model function*. The notation $p(u; \omega|t)$ is used for the value of the conditional probability function of a statistic u given t and under ω.

From the previous discussion it is apparent that if the parametrized family $\mathfrak{P} = \{P_\omega \in \Omega\}$ is a group family under a group H of transformations on the sample $\mathfrak{X}$ then, under mild regularity assumptions, $\mathfrak{X}$ can be viewed as a product space $\mathfrak{U} \times \mathfrak{V}$, the spaces $\mathfrak{V}$, H, and Ω may be identified, and $\mathfrak{P}$ has a model function of the form

$$p(x; \omega) = p(u)p(\omega^{-1}(v)|u)$$

in an obvious notation.

The r-dimensional normal distribution with mean (vector) ξ and variance (matrix) Σ will be indicated by $N_r(\xi, \Sigma)$, and $\mathfrak{N}_r$ will stand for the class of these distributions. (The index r will be suppressed when $r = 1$.) The *precision* (matrix) Δ for $N_r(\xi, \Sigma)$ is the inverse of the variance, i.e. $\Delta = \Sigma^{-1}$. The probability measure of $N_r(\xi, \Sigma)$ will be denoted by $P_{(\xi,\Sigma)}$ or $P_{(\xi,\Delta)}$ according as the parametrization of $\mathfrak{N}_r$ by (ξ, Σ) or by (ξ, Δ) is the one of interest.

The symbol ▶ designates the end of proofs and examples.

1.3 PARAMETRIC MODELS

The statistical models considered in this tract are nearly all parametric and determined by a model function $p(x; \omega)$. Rather more attention than is usual will be given to the parametric aspect of the models, i.e. to the variation domains of the parameters and subparameters involved and to the structure of $p(x: \omega)$ as a function of ω. Thus the observation aspect and the parameter aspect are treated on a fairly equal footing. There are several reasons for this. The most substantial is that the logic of inferential separation cannot be built without certain precise specifications of the role of the parameters. Secondly, it is natural in connection with a comparative discussion of likelihood functions and plausibility functions to give an exposition of Barnard's theory of lods functions, and in a considerable and fundamental portion of that theory observations and parameters occur in a formally equivalent, or completely dual, way. Finally, the stressing of the similarity or duality of the observation and parameter aspects, as far as is statistically meaningful, leads to a certain unification and complementation of the theoretical developments.

There are two main classes of parametric models: the exponential families and the group families. The exponential families, the exact theory of which is a main topic of this book, are determined by model functions of form

$$p(x; \omega) = a(\omega)b(x)\,e^{\theta \cdot t}$$

where θ is a k-dimensional parameter (function) and t is a k-dimensional statistic.

Group families typically have model functions which may be written

$$p(x; \omega) = p(u)p(\omega^{-1}(v)|u),$$

as explained in Section 1.1. A theory of group families—the theory of structural inference—has been developed by Fraser (1968, 1976) (see also Dawid, Stone and Zidek 1973) from Fisher's ideas on fiducial inference. Although the core of fiducial/structural inference is a notion of induced probability distributions for parameters which few persons have found acceptable, the theory comprises many results that are highly useful in the handling of group families along more conventional lines.

The overlap between the two classes of families is very little; thus in the case ω is one-dimensional, the only notable instances of families which belong to both classes appear to be provided by the normal distributions with a known variance and the gamma distributions with a known shape parameter (cf. Lindley 1958, Pfanzagl 1972, and Hipp 1975). Moreover, essential distinctions exist between the mathematical–statistical analyses which are appropriate for each of the two classes. It is remarkable indeed that both classes and the basic difference in their nature were first indicated in a single paper by Fisher (1934).

Each class covers a multitude of important statistical models and allows for a powerful general theory. This strongly motivates studying these classes *per se* and choosing the model for a given data set from one of the two classes, when feasible. Once this is realized, it seems of secondary interest only that one may be led to consider, for instance, exponential families by arguing from various viewpoints of a principled character, such as sufficiency, maximum likelihood, statistical mechanics, etc. (see the references in Section 8.4), especially since each of these viewpoints and its consequences only encompass a fraction of what is of importance in statistics.

PART I

Lods Functions and Inferential Separation

Log-probability functions, log-likelihood functions and log-plausibility functions are the three main instances of lods functions. It is an essential feature of the theory of lods functions that it incorporates a considerable part of the statistically relevant duality relations which exist between the sample aspect and the parameter aspect of statistical models.

Separate inference is inference on a parameter of interest from a part of the original model and data. Margining to a sufficient statistic and conditioning on an ancillary statistic are key procedures for inferential separation.

CHAPTER 2

Likelihood and Plausibility

In this short chapter important basic properties of likelihood functions and plausibility functions are discussed, with particular reference to similarities and differences between these two kinds of function. As a preliminary, the definition and some properties of universality are presented. Universality will also be of significance in the discussions, in subsequent chapters, of prediction, inferential separation, and unimodality.

2.1 UNIVERSALITY

The concept of universality is of significance in the discussions, given later in the book, on plausibility, M-ancillarity, prediction and unimodality.

The probability function $p(\cdot; \omega)$ is said to have a point x as mode point if

$$p(x; \omega) = \sup_x p(x; \omega).$$

and the set of mode points of $p(\cdot; \omega)$ will be denoted by $\check{x}(\omega)$. More generally, x will be called a *mode point for the family* $\mathfrak{p}$ provided that for all $\varepsilon > 0$ there exists an $\omega \in \Omega$ such that

$$(1 + \varepsilon)\, p(x; \omega) \geq \sup_x p(x; \omega). \tag{1}$$

With this designation *universality* of the family $\mathfrak{p}$ is defined as the property that every realizable x is a mode point for $\mathfrak{p}$. If, in fact, every realizable x is a mode point for some member of $\mathfrak{p}$ then $\mathfrak{p}$ is called *strictly universal.*

For convenience in formulation, universality and strict universality will occasionally be spoken of as if they were possible properties of the family of probability measures $\mathfrak{P}$ rather than of $\mathfrak{p}$. Thus, for instance, '$\mathfrak{P}$ is universal' will mean that the family $\mathfrak{p}$ of probability functions determining $\mathfrak{P}$ is universal.

A family $\mathfrak{p}$ for which $\sup_x p(x; \omega)$ is independent of ω will be said to have *constant mode size.*

Most of the standard families of densities are universal, and many examples of universal families will be mentioned later on. Clearly, one has:

Lemma 2.1. *Let H be a class of transformations on $\mathfrak{X}$. Suppose H is transitive and*

that for some $\omega_0 \in \Omega$

$$\mathfrak{p} = \{p(h^{-1}(\cdot); \omega_0): h \in H\}.$$

Then $\mathfrak{p}$ *is universal with constant mode size.*

As a simple consequence of the definition of mode point one finds:

Theorem 2.1. *Let* t *be a statistic and let*

$$p(x; \omega) = p(t; \omega)p(x; \omega|t)$$

be the factorization of the probability function of x *into the marginal probability function for* t *and the conditional probability function for* x *given* t.

Suppose x_0 *is a mode point of* $\mathfrak{p}$ *and set* $t_0 = t(x_0)$. *Then* x_0 *is a mode point of the family of conditional probability functions*

$$\{p(\cdot; \omega|t_0): \omega \in \Omega\}.$$

Corollary 2.1. *If* $\mathfrak{p}$ *is universal then for any given value of* t *the family of conditional probability functions*

$$\{p(\cdot; \omega|t): \omega \in \Omega\}$$

is also universal.

Furthermore, it is trivial that if $\mathfrak{p}$ has only a single member p, say, then $\mathfrak{p}$ is universal if and only if p is constant, i.e. the density is uniform.

The family $\mathfrak{p}$ will be said to *distinguish between the values of* x if for every pair x' and x'' of values of x there exists an $\omega \in \Omega$ such that

$$p(x'; \omega) \neq p(x''; \omega).$$

If $\mathfrak{p}$ is universal and distinguishes between the values of x then, under very mild regularity conditions, x is minimal sufficient. To see this, let x' and x'' be realizable points of $\mathfrak{X}$ and suppose that

$$c'p(x'; \omega) = c''p(x''; \omega) \quad \text{for every } \omega \in \Omega$$

where c' and c'' are constants (which may depend, respectively, on x' and x''). By the universality of $\mathfrak{p}$, the ratio c''/c' must be 1, and this implies $x' = x''$ since $\mathfrak{p}$ distinguishes between values of x. In other words, the partition of $\mathfrak{X}$ induced by the likelihood function is (equivalent to) the full partition into single points; the result now follows from Corollary 4.3.

2.2 LIKELIHOOD FUNCTIONS AND PLAUSIBILITY FUNCTIONS

A brief, comparative discussion of basic properties of likelihood and plausibility functions is given here.

Both likelihood and plausibility functions are considered as determined only

up to a factor which does not depend on the parameter of the model. However, unless otherwise stated, the notations L and Π will stand for the particular choices

$$L(\omega) = L(\omega; x) = p(x; \omega)$$

$$\Pi(\omega) = \Pi(\omega; x) = p(x; \omega)/\sup_x p(x; \omega)$$

of the likelihood and plausibility functions, based on an observation x.

It is important to note that L and Π differ only by a factor

$$s(\omega) = \sup_x p(x; \omega)$$

which is independent of x.

This implies that $\ln \Pi(\omega; x)$, as well as $\ln L(\omega; x)$, is a b-lods function corresponding to the f-lods function $\ln p(x; \omega)$—in the terminology of Barnard's (1949) fundamental theory of lods. Some important common properties of likelihood and plausibility functions may be derived naturally in the theory of lods functions (see Section 3.2).

The *normed* likelihood and plausibility functions will be denoted by $\bar{L}$ and $\bar{\Pi}$, i.e.

$$\bar{L}(\omega) = L(\omega)/\sup_\omega L(\omega)$$

$$\bar{\Pi}(\omega) = \Pi(\omega)/\sup_\omega \Pi(\omega).$$

Clearly,

$$\sup_\omega \Pi(\omega; x) = 1$$

if and only if x is a mode point of $\mathfrak{p}$. Hence, $\Pi = \bar{\Pi}$ for every x if and only if $\mathfrak{p}$ is universal.

For any family $\mathfrak{p}$ such that $\sup_\omega p(x; \omega) < \infty$ for every x one has

$$\bar{L}(\omega; x) = s(\omega) r(x) \bar{\Pi}(\omega; x)$$

where

$$r(x) = \sup_\omega \Pi(\omega; x)/\sup_\omega p(x; \omega).$$

If $\bar{L}$ and $\bar{\Pi}$ are equal for a given value of x then $s(\omega)$ must be independent of ω on the set $\{\omega: p(x; \omega) > 0\}$ which means that $\mathfrak{p}$ has constant mode size on that set. On the other hand, constant mode size of $\mathfrak{p}$ obviously implies that $\bar{L} = \bar{\Pi}$ for every x. In particular, $\bar{L}$ and $\bar{\Pi}$ are thus equal for every x if $\mathfrak{p}$ is generated by a transitive set of transformations.

The set of maximum points of the likelihood or plausibility function constitute respectively the maximum likelihood estimate $\hat{\omega}(x)$ and the maximum plausibility estimate $\check{\omega}(x)$ of the parameter ω, i.e.