Theoretical Statistics

D.R. Cox and D.V. Hinkley

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Theoretical Statistics

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Theoretical Statistics

D.R. Cox
Imperial College, London

D.V. Hinkley
University of Minnesota

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

First issued in hardback 2017

ISBN-13: 978-0-4121-6160-5 (pbk)
ISBN-13: 978-1-1384-6960-0 (hbk)

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site
http://www.crcpress.com

Library of Congress Cataloging-in-Publication Data

Catalog record is available from the Library of Congress.

TO

JOYCE AND BETTY

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

CONTENTS

PREFACE

This book is intended for statisticians wanting a fairly systematic development of the theory of statistics, placing main emphasis on general concepts rather than on mathematical rigour or on the detailed properties of particular techniques. The book is meant also to be suitable as a text-book for a second or third course in theoretical statistics. It is assumed that the reader is familiar with the main elementary statistical techniques and further has some appreciation of the way statistical methods are used in practice. Also knowledge of the elements of probability theory and of the standard special distributions is required, and we assume that the reader has studied separately the theory of the linear model. This is used repeatedly as an illustrative example, however.

We have reluctantly decided to exclude numerical examples and to attempt no detailed discussion of particular advanced techniques or of specific applications. To have covered these would have lengthened and changed the character of the book and to some extent would have broken the thread of the argument. However, in the training of students the working of set numerical exercises, the discussion of real applications and, if at all possible, involvement in current applied statistical work are of very great importance, so that this book is certainly not intended to be the basis of a self-contained course in statistics. To be quite specific, the more discursive parts of the book, for example on the usefulness and limitations of significance tests, will probably not be understandable without some experience of applications.

The mathematical level has been kept as elementary as possible, and many of the arguments are quite informal. The "theorem, proof" style of development has been avoided, and examples play a central role in the discussion. For instance, in the account of asymptotic theory in Chapter 9, we have tried to sketch the main results and to set out their usefulness and limitations. A careful account of the general theorems seems, however, more suited for a monograph than to a general book such as this.

Specialized topics such as time series analysis and multivariate

analysis are mentioned incidentally but are not covered systematically.

A major challenge in writing a book on theoretical statistics is that of keeping a strong link with applications. This is to some extent in conflict with the essential need for idealization and simplification in presentation and it is too much to hope that a satisfactory compromise has been reached in this book. There is some discussion of the connexion between theory and applications in Chapter 1.

The book deals primarily with the theory of statistical methods for the interpretation of scientific and technological data. There are applications, however, where statistical data are used for more or less mechanical decision making, for example in automatic control mechanisms, industrial acceptance sampling and communication systems. An introduction to statistical decision theory is therefore included as a final chapter.

References in the text have been restricted largely to those giving direct clarification or expansion of the discussion. At the end of each chapter a few general references are given. These are intended partly to indicate further reading, partly to give some brief historical background and partly to give some of the main sources of the material in the text. We felt that a very extensive bibliography would be out of place in a general book such as this.

At the end of every chapter except the first, there is an outline of some of the topics and results that it has not been feasible to include in the text. These serve also as exercises, although as such the level of difficulty is very variable and in some cases much detailed work and reference to original sources will be necessary.

D.R. Cox
D.V. Hinkley
London, 1973

1 INTRODUCTION

1.1 Objectives of statistical analysis and theory

Statistical methods of analysis are intended to aid the interpretation of data that are subject to appreciable haphazard variability. The theory of statistics might then be expected to give a comprehensive basis for the analysis of such data, excluding only considerations specific to particular subject matter fields. In fact, however, the great majority of the theory, at any rate as presented in this book, is concerned with the following narrower matter.

There is chosen a family $\mathcal{F}$ of probability models, often completely specified except for a limited number of unknown parameters. From the data under analysis it is required to answer questions of one or both of the following types:

(a) Are the data consistent with the family $\mathcal{F}$?

(b) Assuming provisionally that the data are derived from one of the models in $\mathcal{F}$, what can be concluded about values of unknown parameters, or less commonly about the values of further observations drawn from the same model?

Problems (a) and (b) are related, but the distinction is a useful one. To a very large extent arithmetical rather than graphical methods of analysis are considered.

To illustrate the discussion consider the standard normal-theory model of simple linear regression. According to this, for data consisting of n pairs $(x_1, y_1), \ldots, (x_n, y_n)$, it is supposed that $y_1, \ldots, y_n$ correspond to random variables $Y_1, \ldots, Y_n$ independently normally distributed with constant variance σ^2 and with expected values

$$E(Y_j) = \gamma + \beta x_j \quad (j = 1, \ldots, n), \tag{1}$$

where γ and β are unknown parameters and $x_1, \ldots, x_n$ are regarded as known constants. This is a family $\mathcal{F}$ of models; a particular model is

obtained by specifying values for the parameters γ, β and σ^2. Often, but by no means always, primary interest would be in β.

The problem of type (a) would now be to examine the data for consistency with $\mathcal{F}$, some possibly important types of departure including non-linearity of the dependence of $E(Y_j)$ on x_j, non-constancy of $\text{var}(Y_j)$, lack of independence of the different Y_j's and non-normality of the distribution of the Y_j's. For problems of type (b) it would be assumed provisionally that the data are indeed derived from one of the models in $\mathcal{F}$ and questions such as the following would be considered:

Within what limits is it reasonable to suppose the parameter β to lie?

Are the data consistent with a particular theoretical value for, say, the parameter β?

Let $Y^\dagger$ be a further observation assumed to be obtained from the same model and parameter values as the original data, but with

$$E(Y^\dagger) = \gamma + \beta x^\dagger.$$

Within what limits is $Y^\dagger$ expected to lie?

In the theoretical discussion, it will be usual to take a family of models as given. The task is to formulate meaningful and precise questions about such models and then to develop methods of analysis that will use the data to answer these questions in a sensitive way.

In more detail, in order to deal with the first of these questions just mentioned, the first step is to formulate a precise meaning for such limits for β; this is done in the concept of confidence limits (Section 7.2). Then, there usually being many ways of computing confidence limits, at least approximately, the next step is to define criteria giving the best such limits, or failing that, at least reasonably good confidence limits. Finally, general constructive procedures are required for computing good confidence limits for specific problems. This will lead to sensitive procedures for the analysis of data, assuming that the family $\mathcal{F}$ is well chosen. For example, in the special linear regression model (1), we would be led to a procedure for computing from data limits for, say, the parameter β which will in a certain sense give the most precise analysis of the data given the family $\mathcal{F}$.

In this, and corresponding more general theoretical discussions, much emphasis is placed on finding optimal or near optimal procedures within a family of models.

How does all this correspond with the problems of applied statistics? The following points are relevant:

(i) The choice of the family $\mathcal{F}$ of models is central. It serves first to define the primary quantities of interest, for example, possibly the parameter β in (1), and also any secondary aspects of the system necessary to complete the description. Some of the general considerations involved in the choice of $\mathcal{F}$ are considered in Section 1.2.

(ii) Except occasionally in very simple problems the initial choice of $\mathcal{F}$ will be made after preliminary and graphical analysis of the data. Furthermore, it will often be necessary to proceed iteratively. The results of analysis in terms of a model $\mathcal{F}$ may indicate a different family which may either be more realistic or may enable the conclusions to be expressed more incisively.

(iii) When a plausible and apparently well-fitting family is available it is attractive and sometimes important to use techniques of analysis that are optimal or nearly so, particularly when data are rather limited. However, criteria of optimality must always be viewed critically. Some are very convincing, others much less so; see, for example, Sections 5.2(iii) and 8.2. More importantly, it would be poor to use an analysis optimal for a family $\mathcal{F}$ if under an undetectably different family $\mathcal{F}'$ the same analysis is very inefficient. A procedure that is reasonably good for a broad range of assumptions is usually preferable to one that is optimal under very restrictive assumptions and poor otherwise. Thus it is essential to have techniques not only for obtaining optimal methods but also for assessing the performance of these and other methods of analysis under non-standard conditions.

(iv) At the end of an analysis, it is always wise to consider, even if only briefly and qualitatively, how the general conclusions would be affected by departures from the family $\mathcal{F}$. Often this is conveniently done by asking questions such as: how great would a particular type of departure from $\mathcal{F}$ have to be for the major conclusions of the analysis to be altered?

(v) It is important to adopt formulations such that the statistical analysis bears as directly as possible on the scientific or technological meaning of the data, the relation with previous similar work, the connexion with any theories that may be available and with any practical application that may be contemplated. Nevertheless, often the statistical analysis is just a preliminary to the discussion of the underlying meaning of the data.

Some of these points can be illustrated briefly by the regression

problem outlined above.

The minimal preliminary analysis is a plot of the points (x_j, y_j) on a scatter diagram. This would indicate, for example, whether a transformation of the variables would be wise before analysing in terms of the model (1) and whether there are isolated very discrepant observations whose inclusion or exclusion needs special consideration. After an analysis in terms of (1), residuals, i.e. differences between observed values and estimated values using the model (1), would be calculated, and analysis of these, either graphically or arithmetically, might then suggest a different family of models. With a more complex situation, an initial family of models may be complicated and one might hope to be able to pass to a simpler family, for example one in which there were appreciably fewer unknown parameters. In the present case, however, any change is likely to be in the direction of a more complex model; for instance it may be appropriate to allow $\text{var}(Y_j)$ to be a function of x_j.

It might be argued that by starting with a very complex model, some of the difficulties of (ii)–(iv) could be avoided. Thus one might set up a very general family of models involving additional unknown parameters describing, for example, the transformations of the x- and y-scales that are desirable, the change with x of $\text{var}(Y)$, the lack of independence of different observations and non-normality of the distributions. The best fitting model within this very general family could then be estimated.

While this type of approach is sometimes useful, there are two serious objections to it. First, it would make very heavy going of the analysis even of sets of data of simple structure. A more fundamental reason that a single formulation of very complex models is not in general feasible is the following.

In designing an experiment or scheme of observations, it is important to review beforehand the possible types of observations that can occur, so that difficulties of analysis and ambiguities of interpretation can be avoided by appropriate design. Yet experience suggests that with extensive data there will always be important unanticipated features. An approach to the analysis of data that confines us to questions and a model laid down in advance would be seriously inhibiting. Any family of models is to be regarded as an essentially tentative basis for the analysis of the data.

The whole question of the formulation of families of models in the light of the data is difficult and we shall return to it from time

to time.

To sum up, the problems we shall discuss are closely related to those of applied statistics but it is very important to ensure that the idealization which is inevitable in formulating theory is not allowed to mislead us in practical work.

1.2 Criteria for the choice of families of models

In the previous section the importance has been explained of choosing a family $\mathcal{F}$ of probabilistic models in terms of which questions are to be formulated and methods of statistical analysis derived. For the more elaborate parts of statistical theory we start from the representation of observations in terms of random variables and the idea that normally the parameters of the underlying probability distributions are the quantities of real interest. Yet this view needs to be treated with proper caution and should not be taken for granted. Where the data are obtained by the random sampling of a physically existing population, the parameters have a reasonably clear meaning as properties of the population. In other cases the probability distributions refer to what would happen if the experiment were repeated a large number of times under identical conditions; this is always to some extent hypothetical and with "unique" data, such as economic time series, repetition under identical conditions is entirely hypothetical. Nevertheless the introduction of probability distributions does seem a fruitful way of trying to separate the meaningful from the accidental features of data. Parameters are to be regarded sometimes as representing underlying properties of a random system and sometimes as giving concise descriptions of observations that would be obtained by repetition under the same conditions.

The methods of most practical value are those that combine simple description of the data with efficient assessment of the information available about unknown parameters.

It is hard to lay down precise rules for the choice of the family of models, but we now list briefly some of the considerations involved. These include:

(a) The family should if possible usually establish a link with any theoretical knowledge about the system and with previous experimental work.

(b) There should be consistency with known limiting behaviour. For example, it may be known or strongly suspected that a curve

approaches an asymptote or passes through some fixed point such as the origin. Then, even though this limiting behaviour may be far from the region directly covered by the data, it will often be wise to use a family of models consistent with the limiting behaviour or, at least, to use a family reducing to the limiting behaviour for special parameter values.

(c) So far as possible the models should be put in a form such that the different parameters in the model have individually clear-cut general interpretations.

(d) Further, a description of the data containing few parameters is preferable. This may involve looking at a number of different families to find the most parsimonious representation. There is some chance of conflict with requirements (a) and (b). Indeed in some cases two different analyses may be desirable, one bringing out the link with some relevant theory, the other expressing things in their most economical form.

(e) It is, of course, desirable that the statistical theory associated with the family of models should be as simple as possible.

A fairly recent trend in statistical work places some emphasis on the construction of special models, often by constructing stochastic processes representing the system under investigation. To be at all realistic these often have to be extremely complex. Indeed they may contain so many adjustable parameters that little can be learnt by fitting them to any but extremely extensive data. It is therefore worth stressing that very simple theoretical analyses can be valuable as a basis for statistical analysis. We then choose a family reducing to the theoretical model for special values of certain unknown parameters. In that way we can find how much of the variation present can be accounted for by the simple theory and also characterize the departures from the theory that may be present. It is not at all necessary for a theory to be a perfect fit for it to be a useful tool in analysis.

It will be seen later that the statistical theory of a family is simplified appreciably whenever the family is of the exponential type; see Section 2.2(vi). This provides a fairly flexible set of models and it will often be best to start with a model of this type; for example, nearly all the procedures of multivariate analysis are closely associated with the family of multivariate normal models. The effect of departures can then be considered. It is, however, too much to hope that a reasonably adequate model of exponential type can always be found without violating some other more important requirement.

The requirement of simplicity of statistical analysis may well be in conflict with one or more of the other requirements, for example (a) and (b). Thus the widely used procedure of fitting polynomial curves and surfaces leads to representations whose limiting behaviour is usually unreasonable. This may not matter where the polynomials are used in a limited way for interpolation but seems often likely to be a severe restriction on their usefulness in the interpretation of data; exceptions are when a small quadratic term is used to assess the direction and magnitude of the departure from an expected linear relationship, and when local behaviour in the neighbourhood of a stationary value is under investigation.

1.3 The analysis of complex responses

A widely occurring problem in applied statistics, particularly where automatic recording equipment is used, is the analysis of data in which the response of each individual (subject, animal, batch of material, etc.) is complex and, for example, may consist of a long sequence of observations possibly of several types. Thus in some types of psychological experiment, there will be for each subject a response consisting of a long sequence of successes and failures at some task. In all there may be several such trials for each of many subjects, the whole covering a number of experimental treatments.

It may sometimes be feasible to formulate a single family of models that will embrace all aspects of the data, but more commonly it will be wise to proceed in several steps. First we try to summarize each complex response in a small number of quantities. These may be derived from some formal theoretical analysis, for instance from a simplified mathematical theory of the phenomenon. Or again the formal techniques of time series analysis and multivariate analysis are often guides to the choice of summarizing values. Finally such quantities may be derived from a qualitative inspection and graphical analysis of a sample of individual responses. That is, we construct a few quantities thought to summarize interesting features of the response, such as, for example, average level, trend, amount and nature of local variability. With very complex data it may take much investigation to find the best quantities for further analysis.

Then at the second stage of the analysis the output of the first stage is used as the input data to estimate treatment effects, etc., often by a linear model. There is the following important implication

for the discussion in the rest of this book: the observations for analysis and for which we set up a probability model may be the original data or they may be quantities derived from the preliminary analysis of more complex data. This fact gives the relatively simple models that we shall discuss much wider usefulness than would otherwise be the case.

1.4 Plan of the book

Chapter 2 introduces some fundamental ideas about likelihood and sufficiency which are central to much of the subsequent work. The second part of the chapter, which can be omitted on a first reading, compares some of the broad general approaches to statistical inference.

Chapters 3–7, which in some ways are the core of the book, deal with significance testing and with interval estimation. Chapter 8 is concerned with point estimation. Chapter 9 is on asymptotic theory. For large samples this gives approximate solutions to problems for which no simple "exact" solutions are possible by the techniques of Chapters 3–7.

Finally the last two chapters deal with procedures based on the availability of a prior probability distribution for the unknown parameters of the family $\mathcal{F}$, and with decision theory.

To some extent, the chapters can be taken in a different order. For example, some readers may want to take the rather introductory Chapters 10 and 11 relatively early.

It would be very restricting to use throughout a single completely uniform notation and it is virtually impossible to avoid use of the same letter for different things in different contexts. However, as far as is practicable, the following conventions have been followed.

Random variables representing observations or functions calculated from observations are denoted by capital letters such as X, Y and Z. For example, Y denotes the set of random variables $Y_1, \ldots, Y_n$ considered as a column vector. On the whole Y is reserved for the observation under analysis. To distinguish between random variables and the particular realizations of them forming the data under analysis, the observed values will be denoted by corresponding lower case letters, for example y. It is, however, occasionally convenient to be inconsistent and to retain the capital letters for the observations, where there is no danger of confusion.

Random variables that are not observable, for example "errors", are sometimes denoted by the Greek letters ϵ and η. The Greek letter α is almost exclusively reserved for certain probabilities arising in the study of tests and interval estimation.

Fixed constants are denoted by lower case letters $a, b, c, \ldots$

Unknown parameters are usually denoted by the Greek letters θ, $\phi, \ldots$ They may be scalars or vectors depending on the context; when vectors they are to be considered as column vectors. In dealing with particular situations standard symbols are used; for example, the mean and variance of a normal distribution are denoted by μ and σ^2.

For a random variable, say U, the cumulative distribution function is denoted by $F_U(x)$, or by $F_U(x\,;\theta)$ if it is required to stress dependence on a parameter θ. That is, $F_U(x) = \text{pr}(U \leqslant x)$, where $\text{pr}(A)$ denotes the probability of the event A. The corresponding probability density function, a term used for both continuous and discrete random variables, is denoted by $f_U(x)$ or $f_U(x\,;\theta)$. Thus if U has a Poisson distribution of mean μ, then

$$f_U(x\,;\mu) = \frac{e^{-\mu}\mu^x}{x!} \quad (x = 0, 1, 2, \ldots),$$

whereas if V has a normal distribution of mean μ and variance σ^2, written $N(\mu, \sigma^2)$, then

$$f_V(x\,;\mu, \sigma) = \frac{1}{\sqrt{(2\pi)}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}.$$

The p dimensional multivariate normal distribution of mean the column vector μ and covariance matrix $\mathbf{\Sigma}$ is denoted by $MN_p(\mu, \mathbf{\Sigma})$. A direct extension of the notation for densities is used for joint and conditional densities.

Standard notation is used for expectations, variances and covariances. Where it is required to show the parameter value under which, say, an expectation is calculated, we write, for example, $E(Y\,;\theta)$.

Bold fount is restricted to matrices; the transpose of a matrix or vector is denoted by a superscript T.

An estimate of a parameter θ will sometimes be denoted $\tilde{\theta}$; the notation $\hat{\theta}$ will be restricted to maximum likelihood and least squares estimates. All such estimates $\tilde{\theta}$ are functions of the data, i.e. correspond to random variables. Therefore if the general conventions were to be followed a capital letter would be used for the random variable, but this is often inconvenient. Sums of squares and mean squares

arising in linear model analyses have been denoted SS and MS, sometimes with an appropriate suffix, it being clear from the context whether random variables or observed values are involved.

Asterisks are very largely reserved for tabulated constants arising in significance tests, etc. By convention we use values corresponding to an upper tail area. For example, k_α^* denotes the upper α point of the standard normal distribution, i.e. $\Phi(k_\alpha^*) = 1 - \alpha$, where $\Phi(.)$ is the standard normal integral.

Abbreviations have been kept to a minimum. The only ones widely used are i.i.d. (independent and identically distributed), p.d.f. (probability density function), referring to both continuous and discrete random variables, and m.l.e. (maximum likelihood estimate).

Bibliographic notes

Discussion of the relation between theory and application is most often found in reports of general lectures. Presidential addresses to the Royal Statistical Society by Fisher (1953), Pearson (1956), Kendall (1961), Bartlett (1967), Yates (1968) and Barnard (1972) all in part bear on this issue; see also Kempthorne (1966). Neyman (1960) emphasizes the role of special stochastic models. Tukey and Wilk (1966) have argued in favour of an increased emphasis on graphical methods and on descriptive statistics generally, i.e. on methods which are not based on explicit probabilistic arguments.

Of the many books on statistical methods, that of Daniel and Wood (1971) particularly well illustrates the interplay between the analysis of data and the choice of model. Of those on theory, the later chapters of Kempthorne and Folks (1971) give an introduction to many of the topics of the present book from a viewpoint similar to that taken here. Silvey (1970) gives a concise mathematical introduction. The comprehensive book of Rao (1973) emphasizes distribution theory and the linear model. The three volumes of Kendall and Stuart (1967–69) are particularly valuable in providing an introduction to a wide range of special topics.

A connected account of the history of statistical inference is not available at the time of writing. A collection of papers on various historical topics is edited by Pearson and Kendall (1970); there are a number of biographical articles in *International Encyclopedia of Social Sciences* (Sills, 1968).

2 SOME GENERAL CONCEPTS

2.1 The likelihood

(i) Definition

Let observations $y = (y_1, \ldots, y_n)$ be realised values of random variables $Y = (Y_1, \ldots, Y_n)$ and suppose that an analysis of y is to be based on the provisional assumption that the probability density function (p.d.f.) of Y belongs to some given family $\mathcal{F}$. It is not known, however, which particular p.d.f. in $\mathcal{F}$ is the true one. Any particular p.d.f. $f(y)$ specifies how the density varies across the sample space of possible y values. Often, it is useful to invert this property, and to examine how the density changes at the particular observed value y as we consider different possible functions in $\mathcal{F}$. This results in a comparison between possible densities based on ability to "explain" y, and to emphasise this we define the *likelihood* of $f(.)$ at a particular y by

$$\text{lik}\{f(.); y\} = f(y). \tag{1}$$

Usually it is convenient to work with the natural logarithm, denoted by $l\{f(.); y\}$ and called the *log likelihood*

$$l\{f(.); y\} = \log f(y). \tag{2}$$

In most applications, we consider families $\mathcal{F}$ in which the functional form of the p.d.f. is specified but in which a finite number of unknown parameters $\theta = (\theta_1, \ldots, \theta_q)$ are unknown. Then the p.d.f. of Y for given θ can be written $f(y; \theta)$ or, where desirable, as $f_Y(y; \theta)$. The set of allowable values for θ, denoted by Ω, or sometimes by Ω_θ, is called the *parameter space*. Then the likelihood (1) is a function of θ and we can write, always for $\theta \in \Omega$,

$$\text{lik}(\theta; y) = f(y; \theta), \quad l(\theta; y) = \log f(y; \theta). \tag{3}$$

If it is required to stress the particular random variable for which the likelihood is calculated we add a suffix, as in $\text{lik}_Y(\theta\,;y)$. It is crucial that in these definitions θ is the argument of a function and can take any value in Ω. A very precise notation would distinguish between this use of θ and the particular value that happens to be true, but fortunately this is unnecessary, at least in the present chapter.

Suppose that Y is a continuous vector random variable and that we consider a one-one transformation to a new vector random variable Z with non-vanishing Jacobian $\partial y/\partial z$. Then to any $f_Y(.)$ in $\mathcal{F}$ there corresponds a p.d.f. for Z given by

$$f_Z(z) = f_Y(y)\left|\frac{\partial y}{\partial z}\right|,$$

where z is the transformed value of y. Thus, taking the parametric case for simplicity, we see that the likelihood function based on the observed value z of Z is

$$\text{lik}_Z(\theta\,;z) = \text{lik}_Y(\theta\,;y)\left|\frac{\partial y}{\partial z}\right|. \tag{4}$$

This result suggests that if we are interested in comparing two possible values θ_1 and θ_2 of θ in the light of the data and wish to use the likelihood, it is ratios of likelihood values, rather than, say, differences, that are relevant. For such a comparison cannot reasonably depend on the use of y rather than z.

Very commonly the component random variables $Y_1, \ldots, Y_n$ are mutually independent for all densities in $\mathcal{F}$. Then we can write

$$f_Y(y) = \prod_{j=1}^{n} f_{Y_j}(y_j) = \prod_{j=1}^{n} f_j(y_j),$$

say, and in the parametric case we have for the log likelihood

$$l_Y(\theta\,;y) = \sum_{j=1}^{n} \log f_j(y_j\,;\theta) = \sum_{j=1}^{n} l_j(\theta\,;y_j). \tag{5}$$

When the densities $f_j(y)$ are identical, we unambiguously write $f(y)$.

(ii) Some examples

The following examples give some instances of the calculation of likelihoods and of particular results that will be required later.

Example 2.1. Bernoulli trials. Consider n independent binary

observations, i.e. the jth observation is either a "success", $y_j = 1$, or a "failure", $y_j = 0$, the probability θ of success being the same for all trials ($j = 1, \ldots, n$). The observations $y = (y_1, \ldots, y_n)$ then form a sequence of n ones and zeroes and the probability of any particular sequence is a product of terms θ and $1 - \theta$, there being a θ for every one and a $1 - \theta$ for every zero. Thus

$$\text{lik}_Y(\theta\,; y) = \theta^r(1-\theta)^{n-r}, \tag{6}$$

where $r = \Sigma y_j$ is the number of ones in the observed y. To complete the specification we must give the parameter space which would usually, but not necessarily, be the closed interval $0 \leqslant \theta \leqslant 1$.

Example 2.2. Number of successes in n Bernoulli trials. Suppose that we have exactly the situation of the previous example, except that instead of observing the sequence y we observe only the total number r of successes. We represent this by a random variable R having a binomial distribution. Then

$$\text{lik}_R(\theta\,; r) = \binom{n}{r} \theta^r(1-\theta)^{n-r}. \tag{7}$$

Note that if we are interested in the ratio of likelihoods at say θ_1 and θ_2, then (7) and (6) are equivalent.

Example 2.3. Inverse Bernoulli sampling. Suppose that again we have Bernoulli trials but that new trials continue until a preassigned number r of successes has been obtained and that the total number n of trials necessary to achieve this is observed. Then n is the observed value of a random variable N having a negative binomial distribution and

$$\text{lik}_N(\theta\,; n) = \binom{n-1}{r-1} \theta^r(1-\theta)^{n-r}. \tag{8}$$

Again this is equivalent to (6) and (7) in the sense explained in Example 2.2. Different random systems are, however, involved in the three cases.

Example 2.4. Normal-theory linear model. Suppose that the observations y, considered as an $n \times 1$ column vector, form a realization of the vector random variable Y with $E(Y) = \mathbf{x}\beta$, where $\mathbf{x}$ is a known $n \times q_x$ matrix of rank $q_x \leqslant n$, and β is a $q_x \times 1$ column vector of

unknown parameters. Suppose also that $Y_1, \ldots, Y_n$ are independently normally distributed with unknown variance σ^2. Then

$$\text{lik}_Y(\beta, \sigma^2; y) = (2\pi)^{-\frac{1}{2}n}\sigma^{-n}\exp\left\{-\frac{(y-\mathbf{x}\beta)^{\mathrm{T}}(y-\mathbf{x}\beta)}{2\sigma^2}\right\}. \quad (9)$$

We define the residual sum of squares SS_{res} and least squares estimates $\hat{\beta}$ in the usual way by

$$(\mathbf{x}^{\mathrm{T}}\mathbf{x})\hat{\beta} = \mathbf{x}^{\mathrm{T}}y, \text{SS}_{\text{res}} = (y-\mathbf{x}\hat{\beta})^{\mathrm{T}}(y-\mathbf{x}\hat{\beta}).$$

Then it is easily shown that

$$\text{lik}_Y(\beta, \sigma^2; y) = (2\pi)^{-\frac{1}{2}n}\sigma^{-n}\exp\left\{-\frac{\text{SS}_{\text{res}}}{2\sigma^2} - \frac{(\hat{\beta}-\beta)^{\mathrm{T}}\mathbf{x}^{\mathrm{T}}\mathbf{x}(\hat{\beta}-\beta)}{2\sigma^2}\right\}. \quad (10)$$

Thus the likelihood depends on y only through SS_{res} and $\hat{\beta}$. Note that if $q_x = n$, then $\text{SS}_{\text{res}} = 0$ and the likelihood is unbounded at $\sigma = 0$.

The result (10) covers in particular the special case when $Y_1, \ldots, Y_n$ are i.i.d. in a normal distribution of unknown mean μ and unknown variance σ^2, $N(\mu, \sigma^2)$, say. Then (10) is easily seen to simplify to

$$\text{lik}_Y(\mu, \sigma^2; y) = (2\pi)^{-\frac{1}{2}n}\sigma^{-n}\exp\left\{-\frac{\Sigma(y_j-\bar{y}_\cdot)^2 + n(\bar{y}_\cdot-\mu)^2}{2\sigma^2}\right\}, \quad (11)$$

where $\bar{y}_\cdot = \Sigma y_j/n$.

Example 2.5. Discrete time stochastic process. We can always write the joint probability density of random variables $Y_1, \ldots, Y_n$ in the form

$$f_Y(y) = f_{Y_1}(y_1)f_{Y_2|Y_1}(y_2|y_1)f_{Y_3|Y_2,Y_1}(y_3|y_2, y_1)\ldots$$
$$f_{Y_n|Y_{n-1},\ldots,Y_1}(y_n|y_{n-1}, \ldots, y_1).$$

This is particularly useful for processes developing in time and in particular for Markov processes for which

$$f_{Y_r|Y_{r-1},\ldots,Y_1}(y_r|y_{r-1}, \ldots, y_1) = f_{Y_r|Y_{r-1}}(y_r|y_{r-1}),$$

so that for such processes

$$f_Y(y) = f_{Y_1}(y_1)\prod_{j=2}^{n} f_{Y_j|Y_{j-1}}(y_j|y_{j-1}). \quad (12)$$

In particular consider the two-state Markov chain with transition matrix

$$\begin{bmatrix} \theta_{00} & \theta_{01} \\ \theta_{10} & \theta_{11} \end{bmatrix}.$$

If the state of the system is assumed known at time zero all the terms in (12) are given by one of the θ_{rs}'s, so that

$$\text{lik}(\theta\,; y) = \prod_{r,s} \theta_{rs}^{m_{rs}}, \tag{13}$$

where the elements of the matrix $((m_{rs}))$ give the number of one-step transitions from state r to state s. Note that the result (13) applies directly to all discrete state Markov chains with stationary transition matrices.

Example 2.6. Time-dependent Poisson process. To calculate the likelihood for a stochastic process in continuous time involves in principle a new idea in that we are no longer dealing with a finite number or even with a countable number of random variables. However, if we consider a suitable limiting process a likelihood can usually be calculated. Consider a non-stationary Poisson process of rate $\rho(t)$ observed for the time interval $[0, t_0)$. Let events be observed at times $y_1, \ldots, y_n$. Divide the interval $[0, t_0)$ into a large number m of subintervals each of length h so that $mh = t_0$, and denote these intervals by $[a_j, a_j + h)$ for $j = 1, \ldots, m$. Then by the defining properties of the Poisson process, the subinterval $[a_j, a_j + h)$ contributes a factor $\rho(a_j)h + o(h) = \rho(y_i)h + o(h)$ to the likelihood if $a_j \leqslant y_i < a_j + h$ for some i, whereas if an event did not occur in $[a_j, a_j + h)$ the contribution is $1 - \rho(a_j)h + o(h)$. Probabilities referring to disjoint intervals are independent, so that the likelihood is

$$\prod_{i=1}^{n} \{\rho(y_i)h + o(h)\} \prod_{j}^{*} \{1 - \rho(a_j)h + o(h)\}, \tag{14}$$

where Π^* is the product over all j such that $[a_j, a_j + h)$ contains none of $y_1, \ldots, y_n$. As $h \to 0$ this second product tends to

$$\exp\left\{-\int_0^{t_0} \rho(u)du\right\}.$$

If we omit the factor h^n in (14), the omission corresponding to the conversion from a probability to a probability density, it follows that we may take

$$\text{lik}\{\rho(t); y_1, \ldots, y_n\} = \left\{\prod_{j=1}^{n} \rho(y_j)\right\} \exp\left\{-\int_0^{t_0} \rho(u)du\right\}. \tag{15}$$

In particular if $\rho(t) = \alpha e^{\beta t}$, the likelihood is

$$\text{lik}(\alpha, \beta; y_1, \ldots, y_n) = \exp\left\{n \log \alpha + \beta \Sigma y_j - \frac{\alpha}{\beta}(e^{\beta t_0} - 1)\right\}, \qquad (16)$$

and in the stationary Poisson case with $\beta = 0$

$$\text{lik}(\alpha; y_1, \ldots, y_n) = \alpha^n e^{-\alpha t_0}. \qquad (17)$$

Example 2.7. A likelihood that cannot be specified simply. For most probability models the likelihood can be written down immediately, once the model is properly specified. This is not always the case, however, as the following example shows. Suppose that a series of point events is observed in continuous time and that the model is as follows. Points with integer coordinates are displaced by random amounts that are i.i.d. in $N(0, \sigma^2)$. Only the displaced points are observed; the corresponding order of the originating events is not known. The likelihood can then not be written down in a useful form as a function of σ, especially when σ is large compared with one.

(iii) Mathematical difficulties

There are some mathematical difficulties in a general definition of likelihood for continuous random variables arising from the non-uniqueness of probability density functions. These can be changed on sets of measure zero without changing the probability distribution, and hence likelihood also is in a sense not uniquely defined. However in particular applications there is a "regular" form for the density and it is sensible to define likelihood in terms of this. While a measure-theoretic treatment is possible the fact that all observations are rounded, i.e. essentially discrete, justifies the use of the "regular" version. Indeed in a few cases it will be crucial to remember that continuous distributions are used only as approximations. Similar remarks apply to the likelihood for stochastic processes in continuous time.

(iv) Extended definitions of likelihood

In the definitions (1) and (3) of likelihood, we take the p.d.f. of all the random variables representing the data. Sometimes in dealing with complex problems it is useful to apply a one-one transformation of Y into a new vector random variable which is partitioned

into two parts V and W. Provided that the transformation does not depend on the unknown parameters, and this we assume, we can transform the data y into (v, w).

The function of the unknown parameters obtained by considering the p.d.f. of V at the observed value v, i.e.

$$\text{lik}_V(\theta\,;v) \;=\; f_V(v\,;\theta) \tag{18}$$

is called the *marginal likelihood* for the original problem, based on V. It is the likelihood that we would get if v alone were observed. Again, in some situations we may work with the distribution of W conditionally on $V = v$, i.e. define

$$\text{lik}_{W|V}(\theta\,;w|v) \;=\; f_{W|V}(w|v\,;\theta), \tag{19}$$

which we call a *conditional likelihood* for the original problem.

We consider (18) and (19) only in order to obtain functions simpler than $\text{lik}_Y(\theta\,;y)$. While we could consider (18) and (19) for any convenient V and W, there would in general be a loss of information relevant to θ in so doing. We would like to use (18) or (19) only when, for the particular purpose intended, all or nearly all of the information is retained. Unfortunately it is difficult to express this precisely and for that reason we shall not make extensive direct use of marginal and conditional likelihoods, although they will be implicit in much of the discussion of Chapter 5. An example will illustrate the possible gain of simplicity.

Example 2.8. One-way analysis of variance. Consider data represented by random variables $(Y_{11}, Y_{12}; Y_{21}, Y_{22}; \ldots; Y_{m1}, Y_{m2})$ that are independently normally distributed with variance σ^2 and $E(Y_{jk}) = \mu_j$ $(j = 1, \ldots, m)$. That is, there are m pairs of observations, each pair having a different mean. The following discussion extends immediately if there are more than two observations in each group. The unknown parameter is $\theta = (\mu_1, \ldots, \mu_m, \sigma^2)$, and the likelihood of the full data is

$$\begin{aligned}\text{lik}_Y(\theta\,;y) \;&=\; (2\pi\sigma^2)^{-m}\exp\{-\Sigma\Sigma(y_{jk}-\mu_j)^2/(2\sigma^2)\}\\ &=\; (2\pi\sigma^2)^{-m}\exp\{-\Sigma(\bar{y}_{j.}-\mu_j)^2/\sigma^2\}\exp\{-\Sigma\Sigma(y_{jk}-\bar{y}_{j.})^2/(2\sigma^2)\},\end{aligned} \tag{20}$$

where $\bar{y}_{j.}$ is the mean of the jth pair; note that $\Sigma\Sigma(y_{jk}-\bar{y}_{j.})^2 = \text{SS}_w$ is the usual within-groups sum of squares. It is tempting to conclude that, although σ occurs throughout (20), the information about σ is

largely contained in the final factor. Direct examination of (20) is, however, difficult especially for large m, because it is a function of $m + 1$ parameters.

The situation is clarified somewhat by considering a marginal likelihood. Introduce the orthogonal transformation

$$v_j = (y_{j1} - y_{j2})/\sqrt{2}, \quad w_j = (y_{j1} + y_{j2})/\sqrt{2} \quad (j = 1, \ldots, m).$$

The corresponding random variables are independently normally distributed with variance σ^2 because of the orthogonality of the transformation. Further, $V_1, \ldots, V_m$ are i.i.d. in $N(0, \sigma^2)$ and hence the marginal likelihood based on V is

$$\text{lik}_V(\sigma^2; v) = (2\pi\sigma^2)^{-\frac{1}{2}m} \exp\left(-\frac{\Sigma v_j^2}{2\sigma^2}\right) = (2\pi\sigma^2)^{-\frac{1}{2}m} \exp\left(-\frac{\text{SS}_w}{2\sigma^2}\right). \tag{21}$$

Note that only part of the leading factor in (20) appears in (21). Consideration of the marginal likelihood has replaced a problem with $m + 1$ unknown parameters by a problem with a single parameter. In this particular case, because V and W are independently distributed, (21) can be regarded equally as the conditional likelihood based on V given $W = w$.

If we are concerned solely with σ, and $\mu_1, \ldots, \mu_m$ are unknown and arbitrary, use of (21) is certainly convenient. Is there, however, any loss of information about σ when w is ignored? The distribution of W, a vector of m components, involves all $m + 1$ unknown parameters and it is plausible that, not knowing $\mu_1, \ldots, \mu_m$, we cannot extract any information about σ from w. It is difficult to make this notion precise; the topic is mentioned again in Section 9.2.

2.2 Sufficient statistics

(i) Definition

Suppose that observations $y = (y_1, \ldots, y_n)$ form a realization of a random variable Y and that a family $\mathcal{F}$ of possible distributions is specified. A *statistic* is a function $T = t(Y)$; in general T is a vector. Corresponding to the random variable T is an observed value $t = t(y)$. Note that the distinction between a statistic and an estimate is that the former is not necessarily calculated with the objective of being close to a meaningful parameter.

A statistic S is said to be *sufficient* for the family $\mathcal{F}$ if the conditional density

$$f_{Y|S}(y|s)$$

is the same for all the distributions in $\mathcal{F}$. In a parametric case this means that

$$f_{Y|S}(y|s\,;\theta) \tag{22}$$

does not depend upon θ, $\theta \in \Omega$. For the reason explained in Section 2.1 (iii), possible non-uniqueness of the conditional density for continuous random variables need not worry us.

Note that if S is sufficient, so is any one-one function of S.

Example 2.9. Poisson distribution. Suppose that $\mathcal{F}$ specifies that $Y_1, \ldots, Y_n$ are i.i.d. in a Poisson distribution of mean μ. Then $S = Y_1 + \ldots + Y_n$ is sufficient. To prove this, we calculate the conditional distribution (22). To obtain this, note that

$$f_Y(y\,;\mu) = \prod_{j=1}^{n} \frac{e^{-\mu}\mu^{y_j}}{y_j!} = \frac{e^{-n\mu}\mu^{\Sigma y_j}}{\Pi y_j!},$$

whereas

$$f_S(s\,;\mu) = \frac{e^{-n\mu}(n\mu)^s}{s!}.$$

It follows that

$$f_{Y|S}(y|s\,;\mu) = \begin{cases} \dfrac{(\Sigma y_j)!}{\Pi y_j!} \quad \dfrac{1}{n^{\Sigma y_j}} & (\Sigma y_j = s), \\ 0 & (\Sigma y_j \neq s). \end{cases} \tag{23}$$

Because this does not involve μ, the sufficiency is proved. Note that (23) is the multinomial distribution with $s = \Sigma y_j$ trials and probabilities $(1/n, \ldots, 1/n)$, a result of practical value. The sufficient statistic could equally well be taken as the mean $\Sigma Y_j/n$.

Example 2.10. Uniform distribution. Suppose that $Y_1, \ldots, Y_n$ are i.i.d. in a uniform distribution over $(0, \theta)$, and that $S = \max(Y_j)$. Now

$$f_Y(y\,;\theta) = \begin{cases} 1/\theta^n & (s < \theta), \\ 0 & \text{otherwise}, \end{cases}$$

$$f_S(s\,;\theta) = \begin{cases} ns^{n-1}/\theta^n & (s < \theta), \\ 0 & \text{otherwise}, \end{cases}$$

it being assumed that all the observations are non-negative. Thus

$$f_{Y|S}(y|s\,;\theta) = \begin{cases} \dfrac{1}{ns^{n-1}} & (\max(y_j) = s < \theta), \\ 0 & \text{otherwise}. \end{cases} \tag{24}$$

At first sight (24) does depend on θ. However, the restriction $s < \theta$ in (24) is automatically satisfied for all values of θ for which the conditional distribution is defined.

Example 2.11. Order statistics. Suppose that $Y_1, \ldots, Y_n$ are i.i.d. in a continuous distribution and that $\mathcal{F}$ consists of all densities of continuous random variables. Let S be the set of order statistics $(Y_{(1)}, \ldots, Y_{(n)})$, where $Y_{(1)} \leqslant Y_{(2)} \leqslant \ldots \leqslant Y_{(n)}$. The main distributional properties of order statistics are summarized in Appendix 2. Now

$$f_Y(y) = \prod_{j=1}^{n} f(y_j),$$

$$f_S(s) = \begin{cases} n! \prod_{j=1}^{n} f(s_j) & (s_1 \leqslant s_2 \leqslant \ldots \leqslant s_n), \\ 0 & \text{otherwise}. \end{cases}$$

Thus the conditional density of Y given $S = s$ is

$$f_{Y|S}(y|s) = \begin{cases} 1/n! & \text{if } \{y\} \text{ is a permutation of } \{s\}, \\ 0 & \text{otherwise}. \end{cases} \tag{25}$$

Because this does not depend on the density $f(.)$, the sufficiency of S is proved. Note that (25) expresses nothing more than the obvious fact that, given the ordered values, all permutations of them are equally likely under $\mathcal{F}$.

Example 2.12. The likelihood ratio. Suppose that there are just two possible densities for the vector Y, namely $f_0(y)$ and $f_1(y)$. Let $S = f_1(Y)/f_0(Y)$. That is, two points with the same value of S have the same ratio of likelihoods. We prove the sufficiency of S, taking

the discrete case for simplicity. Let Σ_s^* denote summation over all y such that $f_1(y)/f_0(y) = s$. Then

$$\begin{aligned}\text{pr}\left\{Y = y \,\middle|\, \frac{f_1(Y)}{f_0(Y)} = s\,;\, f_0\right\} &= \frac{\text{pr}(Y = y \cap S = s\,;\, f_0)}{\text{pr}(S = s\,;\, f_0)} \\ &= \frac{f_0(y)}{\Sigma_s^* f_0(y)} = \frac{sf_0(y)}{\Sigma_s^* sf_0(y)} \\ &= \frac{f_1(y)}{\Sigma_s^* f_1(y)} \\ &= \text{pr}\left\{Y = y \,\middle|\, \frac{f_1(Y)}{f_0(Y)} = s\,;\, f_1\right\}.\end{aligned}$$

A similar argument applies if there are more than two possible densities in $\mathcal{F}$. Thus if there are $q + 1$ possible densities $f_0(y), \ldots, f_q(y)$, the set of likelihood ratios

$$f_1(y)/f_0(y), \ldots, f_q(y)/f_0(y)$$

is sufficient. The choice of $f_0(y)$ as a reference point is arbitrary and there is an obvious modification if $f_0(y) = 0$. In the parametric case the corresponding result is that for a fixed θ_0 the set of all values

$$f(y\,;\theta)/f(y\,;\theta_0) = \text{lik}(\theta\,;y)/\text{lik}(\theta_0\,;y) \tag{26}$$

for $\theta \in \Omega$ forms a sufficient statistic. That is, if two points y_1 and y_2 have proportional likelihood functions, they have the same value of a sufficient statistic. This result is pursued in part (iv) of this section, in connexion with minimal sufficiency.

(ii) Factorization theorem

The examples of the previous section illustrate the idea of a sufficient statistic but do not show how to find the sufficient statistic in any particular case. Example 2.12 does, however, suggest that it is the structure of the likelihood function that indicates the form of the sufficient statistic. This is expressed formally in the factorization theorem:

A necessary and sufficient condition that S be sufficient for θ in the family $\mathcal{F}$ is that there exist functions $m_1(s, \theta)$ and $m_2(y)$ such that for all $\theta \in \Omega$,

$$\text{lik}(\theta\,;y) = m_1(s, \theta)m_2(y). \tag{27}$$

If S is sufficient, then $f_{Y|S}(y|s)$ does not depend on θ and can be written as $m_2(y)$. Thus

$$\begin{aligned} \text{lik}(\theta\,;y) &= f_Y(y\,;\theta) = f_{Y|S}(y|s)f_S(s\,;\theta) \\ &= m_1(s,\theta)m_2(y), \end{aligned}$$

say. Conversely if (27) holds, we may calculate $f_{Y|S}(y|s\,;\theta)$ as

$$\frac{f_Y(y\,;\theta)}{f_S(s\,;\theta)} = \begin{cases} f_Y(y\,;\theta)/\sum_{z\,:\,s(z)=s} f_Y(z\,;\theta) \text{ in discrete case,} \\ f_Y(y\,;\theta)/\int\ldots\int f_Y(z\,;\theta)\,|J|\,dz \text{ in continuous case.} \end{cases} \tag{28}$$

In the second formula we have changed to new variables (s, z), introduced a Jacobian, J, and integrated with respect to z. If we substitute (27) into (28), the term $m_1(s, \theta)$ cancels and the conditional distribution thus does not involve θ.

Note that in (27) we can, if we wish, arrange that $m_1(s, \theta)$ is the p.d.f. of S. We call (27) the *factorization theorem*; it will be used repeatedly later. Two examples suffice for now.

Example 2.13. Poisson distribution (ctd). For the situation of Example 2.9, where $Y_1, \ldots, Y_n$ are i.i.d. in a Poisson distribution of mean μ, the joint probability is

$$\prod_{j=1}^{n} \frac{e^{-\mu}\mu^{y_j}}{y_j!} = (e^{-n\mu}\mu^{\Sigma y_j}) \cdot \left(\prod \frac{1}{y_j!}\right).$$

We may take the two factors in this as respectively $m_1(s, \mu)$ and $m_2(y)$, so that $S = \Sigma Y_j$ is sufficient.

Example 2.14. Cauchy distribution. If $Y_1, \ldots, Y_n$ are i.i.d. in a Cauchy distribution of location parameter θ, the joint density is

$$\frac{1}{\pi^n} \prod_{j=1}^{n} \frac{1}{\{1 + (y_j - \theta)^2\}}$$

and no factorization involving a function s of fewer than n dimensions is possible; see Example 2.16 for proof. By the general result of Example 2.11 the order statistics of the sample are sufficient.

(iii) Interpretation of sufficiency

Consider two individuals both involved in observations associated

with the family $\mathcal{F}$, as follows:

Individual I observes y, a value of the random variable Y;

Individual II proceeds in two stages:

(a) he observes s, a value of the random variable S having the p.d.f. $f_S(s\,;\theta)$,

(b) he then observes y, a value of a random variable having the p.d.f. $f_{Y|S}(y|s)$, not depending on θ.

The following two statements are very plausible.

(1) Because the final distributions of Y for the two individuals are identical, the conclusions to be reached from a given y are identical for the two individuals.

(2) Because Individual II, in stage (b), is sampling a fixed distribution, i.e. is in effect drawing values from a table of random numbers, only stage (a) is informative, so long as the correctness of $\mathcal{F}$ is postulated.

If both (1) and (2) are accepted, it follows that if y is observed then the conclusions to be drawn about θ depend only on $s = s(y)$, so long as $\mathcal{F}$ is the basis of the analysis.

We shall discuss this further later. The argument can be looked at in two slightly different ways. On the one hand the argument can be thought convincing enough to make it a basic principle that, so long as the correctness of $\mathcal{F}$ is accepted, the conclusions to be drawn should depend only on s. Alternatively one may simply note that all the optimality criteria that we shall consider later lead to the use of s and we may regard the above argument as an explanation of that.

Note that although restriction to the use of s may achieve a big reduction in dimensionality we still have to decide what to do with s, or how to interpret it.

(iv) Minimal sufficiency

If in a particular problem S is a sufficient statistic for θ, then so too is (S, T) for any statistic $T = t(Y)$. Of course, we would rather deal with S than with (S, T) since our object is to summarize the data concisely. If no further reduction from S while retaining sufficiency is possible, then S is said to be *minimal sufficient*; S is necessarily a function of all other sufficient statistics that can be constructed.

Example 2.15. Binomial distribution. Let $Y_1, \ldots, Y_n$ be independent Bernoulli random variables with parameter θ and $S = \Sigma Y_j$. Then $V = g(S)$ is a summary or simplification of S only if $g(r_1) = g(r_2) = v$

for some v and $0 \leqslant r_1 \neq r_2 \leqslant n$. But for $s = r_1, r_2$

$$\text{pr}(S = s \mid V = v) = \frac{\text{pr}(S = s \cap V = v)}{\text{pr}(V = v)}$$

$$= \frac{\binom{n}{s}\theta^s(1-\theta)^{n-s}}{\binom{n}{r_1}\theta^{r_1}(1-\theta)^{n-r_1} + \binom{n}{r_2}\theta^{r_2}(1-\theta)^{n-r_2}},$$

which depends on θ. Thus V is not sufficient and S is minimal sufficient.

We want to use minimal sufficient statistics wherever possible. Sometimes the appropriate factorization of the likelihood is obvious on inspection, particularly for a single parameter. In other cases we can use an important close relationship between minimal sufficiency and ratios of likelihoods to derive the minimal sufficient statistic.

Any statistic, and therefore in particular any sufficient statistic S, divides the sample space into equivalence classes, each class containing all possible observations y with a common value of s. The fact that if S is minimal sufficient so too is any one-one function of S indicates that it is the set of equivalence classes that determines the essential nature of the reduction by minimal sufficiency, rather than the particular labelling of the equivalence classes.

Consider the partition created by putting all points with proportional likelihood functions into the same equivalence class, i.e. define the classes

$$\mathcal{D}(y) = \left\{ z ; \frac{f_Y(z;\theta)}{f_Y(y;\theta)} = h(z,y), \text{ for all } \theta \in \Omega \right\}; \qquad (29)$$

if $z \in \mathcal{D}(y_1)$ and $\mathcal{D}(y_2)$, then $\mathcal{D}(y_1) = \mathcal{D}(y_2)$. This partitioning is minimal sufficient. To see that it is sufficient, note that the conditional distribution of Y within its equivalence class is independent of θ. To show that it is minimal sufficient, consider any other sufficient statistic $V = v(Y)$ which, by the factorization theorem (27), is such that

$$f_Y(y;\theta) = m_1^\dagger\{v(y), \theta\} m_2^\dagger(y) = f_Y(z;\theta)\frac{m_2^\dagger(y)}{m_2^\dagger(z)},$$

if y and z are such that $v(y) = v(z)$. But this implies that y and z are

equivalent in the sense of (29). Therefore the partition (29) includes that based on V, proving the minimal sufficiency of (29).

Thus we inspect the likelihood ratio $f_Y(z\,;\theta)/f_Y(y\,;\theta)$ in order to find which y and z should be assigned the same value of the minimal sufficient statistic.

Example 2.16. Cauchy distribution (ctd). If $Y_1, \ldots, Y_n$ are i.i.d. in the Cauchy distribution of Example 2.14, the likelihood ratio is

$$\frac{f_Y(z\,;\theta)}{f_Y(y\,;\theta)} = \frac{\Pi\{1+(y_j-\theta)^2\}}{\Pi\{1+(z_j-\theta)^2\}}$$

and is thus a rational function of θ. For the ratio to be independent of θ, all powers of θ must have identical coefficients in numerator and denominator. This happens if and only if $(y_1, \ldots, y_n)$ is a permutation of $(z_1, \ldots, z_n)$. Therefore the minimal sufficient statistic is the set of order statistics $(Y_{(1)}, \ldots, Y_{(n)})$.

It was shown in Example 2.11 that the order statistics are sufficient for the full family of continuous distributions; it follows from their minimal property for the Cauchy distribution that *a fortiori* they are minimal for the larger family.

From now on, by sufficient statistic we always mean minimal sufficient statistic.

(v) Examples

We now consider three further examples which serve both to illustrate the factorization theorem and to give some results of intrinsic importance.

Example 2.17. Normal-theory linear model (ctd). The likelihood for the normal-theory linear model was calculated in (10) and involves the observations only through $(\hat{\beta}, \mathrm{SS}_{\mathrm{res}})$. These are therefore sufficient statistics for the unknown parameters (β, σ^2). If, however, the variance is known and equal to say σ_0^2 we can in (10) separate off the factor

$$\exp\left(-\frac{\mathrm{SS}_{\mathrm{res}}}{2\sigma_0^2}\right)$$

and treat it as the function $m_2(y)$ in the factorization theorem (27). It then follows that $\hat{\beta}$ is sufficient for β. In particular, when the

random variables are i.i.d., the sample mean and estimate of variance are sufficient and when the variance is known the mean is sufficient.

Example 2.18. Uniform distribution of zero mean. Suppose that $Y_1, \ldots, Y_n$ are i.i.d. with uniform density over $(-\theta, \theta)$. Then the likelihood is

$$\begin{cases} \left(\frac{1}{2\theta}\right)^n & (\max|y_j| \leqslant \theta), \\ 0 & (\max|y_j| > \theta). \end{cases} \tag{30}$$

Hence the sufficient statistic is $\max|Y_j|$, or equivalently $\max(-Y_{(1)}, Y_{(n)})$, where $Y_{(1)} = \min(Y_j)$, $Y_{(n)} = \max(Y_j)$ are the extreme order statistics.

This may be compared with the rather simpler result of Example 2.10 that for the uniform distribution over $(0, \theta)$ the largest value is sufficient.

A simple extension of (30) shows that for the uniform distribution with both terminals unknown, the smallest and largest values are together sufficient. The same sufficient statistics apply for a uniform distribution of known range but unknown mean, e.g. the uniform distribution from $\theta - 1$ to $\theta + 1$.

These results generalize immediately to a known distribution truncated at unknown points, i.e. to the density

$$\frac{p(y)}{\int_{\theta_1}^{\theta_2} p(u)du} \quad (\theta_1 \leqslant y \leqslant \theta_2), \tag{31}$$

where $p(.)$ is a known non-negative function and one or both of θ_1 and θ_2 are unknown parameters. Again the relevant extreme order statistics are sufficient.

Example 2.19. Life-testing with an exponential distribution of life. Suppose that in observations on n individuals, r "die" after times $y_1, \ldots, y_r$, whereas the remaining $m = n - r$ are still "alive" after times under test of $y_1', \ldots, y_m'$; there are a number of situations in life-testing where data have to be analysed with an appreciable number of lives incomplete. If completed lives are represented by random variables $Y_1, \ldots, Y_n$ which are i.i.d. with p.d.f. $\rho e^{-\rho y}$ $(y \geqslant 0)$, the likelihood is

$$\prod_{j=1}^{r} \rho e^{-\rho y_j} \prod_{k=1}^{m} e^{-\rho y_k'}, \tag{32}$$

the terms for the incomplete lives being the probabilities of times to death exceeding $y_1', \ldots, y_m'$. Thus the likelihood is

$$\rho^r e^{-\rho y_\cdot},$$

where $y_\cdot = \Sigma y_j + \Sigma y_k'$ is the total time at risk; the sufficient statistic is $(R, Y_\cdot)$. This result is the continuous time analogue of the result that in any set of Bernoulli trials the number of successes and the total number of trials form the sufficient statistic.

(vi) Exponential family of distributions

Suppose first that there is a single parameter and that $Y_1, \ldots, Y_n$ are mutually independent with

$$f_{Y_j}(y\,;\theta) = \exp\{a(\theta)b_j(y) + c_j(\theta) + d_j(y)\}, \tag{33}$$

where $a(.)$, $b_j(.)$, $c_j(.)$, $d_j(.)$ are known functions. Then

$$f_Y(y\,;\theta) = \exp\{a(\theta)\Sigma b_j(y_j) + \Sigma c_j(\theta) + \Sigma d_j(y_j)\}, \tag{34}$$

so that $\Sigma b_j(Y_j)$ is sufficient. In particular, if the Y_j are identically distributed, $\Sigma b(Y_j)$ is sufficient. Several interesting special distributions have the form

$$\exp\{a(\theta)b(y) + c(\theta) + d(y)\}, \tag{35}$$

among them the normal, gamma, binomial and Poisson distributions. For example, to see that the gamma distribution with known index k_0 belongs to the family we write

$$\begin{aligned} f_Y(y\,;\rho) &= \rho(\rho y)^{k_0-1}e^{-\rho y}/\Gamma(k_0) \\ &= \exp\{-\rho y + k_0 \log\rho + (k_0 - 1)\log y - \log\Gamma(k_0)\}. \end{aligned}$$

Thus $a(\rho) = -\rho$, $b(y) = y$, $c(\rho) = k_0 \log\rho - \log\Gamma(k_0)$ and $d(y) = (k_0 - 1)\log y$, and the sufficient statistic for ρ from i.i.d. random variables $Y_1, \ldots, Y_n$ is $\Sigma b(Y_j) = \Sigma Y_j$.

One-one transformations of variable or parameter do not affect the general form of (35), i.e. whether or not a distribution belongs to the simple exponential family. Thus we can, provided that $a(.)$ and $b(.)$ are monotonic, transform to a new parameter $\phi = -a(\theta)$ and a new variable $Z = b(Y)$. The p.d.f. for Z has the simple form

$$f_Z(z\,;\phi) = \exp\{-z\phi + c^\dagger(\phi) + d^\dagger(z)\}, \tag{36}$$

and the sufficient statistic for ϕ based on i.i.d. random variables $Z_1, \ldots, Z_n$ is ΣZ_j. The new parameter ϕ is often called the *natural parameter* for the problem, for several technical and practical reasons. One of these is that the ratio of likelihood functions at ϕ_1 and $\phi_2 < \phi_1$ is an increasing function of the sufficient statistic. Also it will turn out that comparisons of different sets of data are most easily achieved in terms of comparisons of the natural parameter values. For example, the natural parameter for the binomial distribution is $\phi = \log\{\theta/(1-\theta)\}$, the so-called log odds ratio. The theory of the comparison of two binomial distributions is simplest not in terms of $\theta_1 - \theta_2$ but in terms of $\phi_1 - \phi_2$. Whether this is really the best parametrization in terms of which to make the comparison depends in addition on circumstances other than mathematical simplicity.

Example 2.20. Exponential family linear model. Suppose that the Y_j's independently have p.d.f.'s $f_{Y_j}(y_j\,;\theta_j)$ belonging to the same exponential family but with different parameter values θ_j, all of which are themselves functions of a single parameter ψ; that is, $\theta_j = \theta_j(\psi)$. The joint distribution in the simplified form (36) is

$$f_Z(z\,;\phi_1, \ldots, \phi_n) = \exp\left\{-\sum_{j=1}^n z_j\phi_j + \sum_{j=1}^n c^\dagger(\phi_j) + \sum_{j=1}^n d^\dagger(z_j)\right\}.$$

Thus if the dependence of θ_j on ψ implies a linear relationship, $\phi_j = a_j\psi$, with a_j constant, then $\Sigma a_j Z_j$ is sufficient for ψ. Linearity on any other scale will not produce a single sufficient statistic for ψ.

The generalization of the exponential family to vector parameters is to consider the density for Y_j given $\theta = (\theta_1, \ldots, \theta_q)$ as

$$f_{Y_j}(y_j\,;\theta) = \exp\left\{\sum_{k=1}^m a_k(\theta)b_{jk}(y_j) + c_j(\theta) + d_j(y_j)\right\}. \tag{37}$$

Then the joint p.d.f. for independent variables $Y_1, \ldots, Y_n$ can be written as

$$f_Y(y\,;\theta) = \exp\left\{\sum_{k=1}^m a_k(\theta)s_k(y) + c.(\theta) + d.(y)\right\} \tag{38}$$

where

$$s_k(y) = \sum_{j=1}^n b_{jk}(y_j) \quad (k = 1, \ldots, m).$$

The sufficient statistic for θ is therefore $S = (S_1, \ldots, S_m)$. The dimensions m and q of s and θ respectively are not necessarily equal. The case $m < q$ might occur if some non-linear relationship exists between the components of θ, but usually this will not happen. Most common is the case $m = q$, which arises in Example 2.4, the normal-theory linear model of full rank. There the dimensionality of $\theta = (\beta, \sigma^2)$ is $q = q_x + 1$ and it follows directly from (10) that $m = q$, the sufficient statistic being $(\hat{\beta}, \text{SS}_{\text{res}})$.

The case $m > q$, while not very common in applications, can arise in a perfectly natural way. We give one example.

Example 2.21. Normal distribution with known coefficient of variation. Consider a normal distribution in which the ratio γ_0 of standard deviation to mean is known, $\sigma = \gamma_0 \mu$, say. The p.d.f. is

$$\frac{1}{\gamma_0 \mu \sqrt{(2\pi)}} \exp\left\{-\frac{(y-\mu)^2}{2\gamma_0^2 \mu^2}\right\}$$

$$= \exp\left\{-\frac{y^2}{2\gamma_0^2 \mu^2} + \frac{y}{\gamma_0^2 \mu} - \frac{1}{2\gamma_0^2} - \frac{1}{2}\log(2\pi\gamma_0^2 \mu^2)\right\}$$

which is of the form (37) with $m = 2$. If $Y_1, \ldots, Y_n$ are i.i.d. with this distribution, the sufficient statistic is $(\Sigma Y_j, \Sigma Y_j^2)$ or equivalently $(\bar{Y}_\cdot, \text{MS})$ with $\bar{Y}_\cdot = \Sigma Y_j/n$ and $\text{MS} = \Sigma(Y_j - \bar{Y}_\cdot)^2/(n-1)$.

Similar examples can be formed from other distributions. A rather different situation is illustrated by Example 2.19, with censored data from an exponential distribution. Here again $q = 1, m = 2$, with the sufficient statistic being number of uncensored observations and the total time on test.

When $q = m$, the p.d.f. (37) can, by transformation of variables and parameters, be taken in a simple form somewhat analogous to (36), namely

$$f_Z(z\,;\phi) = \exp\left\{-\sum_{r=1}^{q} \phi_r b_r^\dagger(z) + c^\dagger(\phi) + d^\dagger(z)\right\}.$$

It is then appealing to assign components of the sufficient statistic

$$S = \left\{\sum_{j=1}^{n} b_1^\dagger(Z_j), \ldots, \sum_{j=1}^{n} b_q^\dagger(Z_j)\right\}$$

to the corresponding components of ϕ. This will be done implicitly

in Section 5.2, in connexion with testing hypotheses about components of θ, but for the moment we keep to the original definition where the whole sufficient statistic is a vector associated with the whole parameter as a vector.

In general, the dimension of the sufficient statistic will not be smaller than the sample size unless the distribution is a member of the exponential family. This is illustrated by Example 2.14, the Cauchy distribution, by the Weibull distribution of unknown index, and by Example 2.11, showing that when $\mathcal{F}$ is the family of all continuous distributions, the order statistics are sufficient; it is easily shown in this last example that no further reduction is possible. We do not here attempt to prove the equivalence of sufficient reduction and the exponential family, but for the one-dimensional case ($m = q = 1$) Exercise 2.11 outlines a method of establishing the connexion.

To emphasise that sufficiency is a property of the sampling model, as well as of the distributions being sampled, we give the following example.

Example 2.22. Change-point model. Suppose that $u_1, \ldots, u_n$ are fixed constants, all different, and that for some unknown ξ, Y_j has the distribution $N(0, 1)$ if $u_j < \xi$ and the distribution $N(\mu, 1)$ if $u_j \geqslant \xi$. Then the sufficient statistic for the unknown parameter $\theta = (\mu, \xi)$ is the full set $Y_1, \ldots, Y_n$, and no reduction is possible.

(vii) Completeness

A mathematically important idea is that of completeness. If S is a sufficient statistic for θ in the family of distributions indexed by $\theta \in \Omega$, then S is called *complete* if a necessary condition for

$$E\{h(S); \theta\} = 0 \quad (\theta \in \Omega) \tag{39}$$

is

$$h(S) = 0 \quad (\theta \in \Omega),$$

except possibly on sets of measure zero with respect to all the distributions concerned. A weaker concept is that of *bounded completeness*, for which $h(S)$ must be bounded. The property of completeness guarantees uniqueness of certain statistical procedures based on S; this we discuss in later chapters.

Example 2.23. Normal mean. Let $Y_1, \ldots, Y_n$ be i.i.d. in $N(\mu, 1)$, and

let $S = \bar{Y}_\cdot = \Sigma Y_j/n$. Then, because $\bar{Y}_\cdot$ is $N(\mu, 1/n)$, the identity (39) becomes

$$\int_{-\infty}^{\infty} h(s)e^{-\frac{1}{2}s^2} e^{ns\mu} ds = 0 \quad (-\infty < \mu < \infty).$$

But the integral is a bilateral Laplace transform, so that by the appropriate inversion theorem we deduce that $h(s)\exp(-\frac{1}{2}ns^2)$ is identically zero except on sets of Lebesgue measure zero. Thus $h(s) \equiv 0$, and S is complete.

Example 2.24. Binomial distribution (ctd). Let $Y_1, \ldots, Y_n$ be independent Bernoulli random variables with parameter θ. Then $S = \Sigma Y_j$ is sufficient for θ. Values of $h(s)$ other than for $s = 0, \ldots, n$ have zero probability and are of no concern; let $h(s) = h_s$. The identity (39) becomes

$$\sum_{s=0}^{n} h_s \binom{n}{s} \theta^s (1-\theta)^{n-s} = 0 \quad (0 \leqslant \theta \leqslant 1). \tag{40}$$

Here the sum is a polynomial of degree n which is identically zero, which implies $h_1 = \ldots = h_n = 0$, so that again S is complete. In fact the vanishing of the h_s's follows if (40) holds for at least $n + 1$ distinct values of θ. That is, S is complete for much smaller parameter spaces than $[0, 1]$.

It can be shown (Lehmann, 1959, Section 4.3) that for random variables i.i.d. in the exponential family density (38), $\dim(S) = \dim(\theta)$, i.e. $m = q$, is necessary and sufficient for S to be complete. Thus in the situation of Example 2.21, the normal distribution with known coefficient of variation, with $m > q$, the sufficient statistic $(\Sigma Y_j, \Sigma Y_j^2)$ is not complete; this is easily verified directly because

$$\frac{n + \gamma_0^2}{1 + \gamma_0^2} \Sigma Y_j^2 - (\Sigma Y_j)^2$$

has expectation zero for all μ. In general if a sufficient statistic is boundedly complete it is minimal sufficient (Lehmann and Scheffé, 1950, 1955); the converse is false.

(viii) Ancillary statistics

Consider the situation where S is the minimal sufficient statistic for θ and $\dim(S) > \dim(\theta)$. Then it sometimes happens that we can

write $S = (T, C)$, where C has a marginal distribution not depending on θ. If this is so, C is called an *ancillary statistic*. Some writers then refer to T as *conditionally sufficient*, because T is used as a sufficient statistic in inference conditionally on $C = c$. The ancillary statistic C is chosen to have maximum dimension.

Example 2.25. Random sample size. Let N be a random variable with a known distribution $p_n = \text{pr}(N = n)$ $(n = 1, 2, \ldots)$, and let $Y_1, \ldots, Y_N$ be i.i.d. with the exponential family density (35). Then the likelihood of the data $(n, y_1, \ldots, y_n)$ is

$$f_{N,Y}(n, y) = p_n \exp\left\{a(\theta) \sum_{j=1}^{n} b(y_j) + nc(\theta) + \sum_{j=1}^{n} d(y_j)\right\}.$$

Thus

$$\left\{\sum_{j=1}^{N} b(Y_j), N\right\}$$

is sufficient for θ, and N is an ancillary statistic, whereas $\Sigma b(Y_j)$ is only conditionally sufficient. Any sample size not fixed in advance, but with known distribution independent of θ, is an ancillary statistic.

Example 2.26. Mixture of normal distributions. Suppose that a random variable Y is equally likely to be $N(\mu, \sigma_1^2)$ or $N(\mu, \sigma_2^2)$, where σ_1 and σ_2 are different and known. An indicator random variable C is observed, taking the value 1 or 2 according to whether Y has the first or the second distribution. Thus it is known from which distribution y comes. Then the likelihood of the data (c, y) is

$$f_{C,Y}(c, y) = \tfrac{1}{2}(2\pi\sigma_c^2)^{-\frac{1}{2}} \exp\{-(y - \mu)^2/(2\sigma_c^2)\},$$

so that $S = (C, Y)$ is sufficient for μ with σ_1^2 and σ_2^2 known. Because $\text{pr}(C = 1) = \text{pr}(C = 2) = \frac{1}{2}$ independent of μ, C is ancillary.

Example 2.27. Normal-theory linear model (ctd). In a linear regression problem, suppose that the values of the explanatory variable have a known joint p.d.f. $f_X(x)$, and that, conditionally on $X = x$, the $Y_1, \ldots, Y_n$ are independent, Y_j having the distribution $N(\gamma + \beta x_j, \sigma^2)$. Then the full likelihood of the data is

$$f_{X,Y}(x, y) = f_X(x)(2\pi\sigma^2)^{-\frac{1}{2}n} \exp\left\{-\frac{1}{2\sigma^2} \sum_{j=1}^{n} (y_j - \gamma - \beta x_j)^2\right\}.$$

The sufficient statistic for $(\gamma, \beta, \sigma^2)$ is

$$S = (\hat{\gamma}, \hat{\beta}, \mathrm{SS}_{\mathrm{res}}, \Sigma X_j, \Sigma X_j^2), \tag{41}$$

the last two components of which form an ancillary statistic. That is, even if the explanatory variable is random, conditioning on the ancillary statistic would lead to treating the explanatory variable as fixed. The argument extends immediately to the general normal-theory linear model.

These simple examples are intended to suggest that inference about θ should be conditional on the ancillary statistic. We can regard the observed value c as describing that part of the total sample space relevant to the problem at hand. For instance, in Example 2.26, the ancillary statistic tells us which normal distribution was in fact applicable. The fact that some other normal distribution might have been used, but actually was not, seems irrelevant to the interpretation of y. However some difficulties are encountered with ancillary statistics. First, there is no general method for constructing C. Secondly, C may not be unique. The following example, given by Basu (1964) in an extended discussion of ancillarity, illustrates both difficulties.

Example 2.28. Special multinomial distribution. Let $Y_1, \ldots, Y_n$ be i.i.d. with the discrete distribution

$$\mathrm{pr}(Y_j = 1) = \tfrac{1}{6}(1-\theta), \quad \mathrm{pr}(Y_j = 2) = \tfrac{1}{6}(1+\theta),$$

$$\mathrm{pr}(Y_j = 3) = \tfrac{1}{6}(2-\theta), \quad \mathrm{pr}(Y_j = 4) = \tfrac{1}{6}(2+\theta).$$

If n_l is the number of observations equal to $l (l = 1, \ldots, 4)$, then the joint probability of a particular sequence is

$$f_Y(y\,;\theta) = 6^{-n}(1-\theta)^{n_1}(1+\theta)^{n_2}(2-\theta)^{n_3}(2+\theta)^{n_4}. \tag{42}$$

The statistic $S = (N_1, N_2, N_3, N_4)$ is minimal sufficient; of course one component can be omitted in view of the identity $\Sigma N_l = n$. The particular structure here leads to two possible ancillary statistics, namely $C_1 = (N_1 + N_2, N_3 + N_4)$ and $C_2 = (N_1 + N_4, N_2 + N_3)$. Which of these to use in inference about θ depends on which one best separates all possible sets of data into equally informative sets; see also Example 2.37 and Exercise 4.11.

While such non-uniqueness arises rather rarely in applications, the possibility is theoretically disturbing.

The next example indicates a very general set of ancillary statistics.

Example 2.29. Location family. Let $Y_1, \ldots, Y_n$ be i.i.d. in the location family with density $h(y-\theta)$. The order statistics $Y_{(1)}, \ldots, Y_{(n)}$ are sufficient by the arguments of Example 2.11. Except when $\log h(y)$ is a polynomial in y of degree less than n, the order statistics are minimal sufficient. The contrasts between these, as determined, for example, by $C_2 = Y_{(2)} - Y_{(1)}$, $C_3 = Y_{(3)} - Y_{(1)}, \ldots, C_n = Y_{(n)} - Y_{(1)}$, are distributed independently of θ and hence form an ancillary statistic. The remaining component of the sufficient statistic can be taken as $T = Y_{(1)}$, or equivalently as any function of T and C, such as $\bar{Y}_.$. The consequences of studying the conditional distribution of T given the ancillary statistic will be taken up in Example 4.15. The statistic $C = (C_2, \ldots, C_n)$ is called the *configuration.*

It would have been possible to have defined ancillary statistics without the preliminary reduction by minimal sufficiency. However, the fact that inference is wherever possible carried out in terms of the minimal sufficient statistic makes the present definition appealing. The alternative would be to define an ancillary statistic as any function of Y with a distribution independent of θ. That this would lead to additional complications is shown by the following example.

Example 2.30. Bivariate normal distribution with unknown correlation coefficient. Suppose that $(Y_1, Z_1), \ldots, (Y_n, Z_n)$ are i.i.d. in a bivariate normal distribution with zero means, unit variances and unknown correlation coefficient ρ. The joint p.d.f. is

$$\frac{1}{(2\pi)^n (1-\rho^2)^{\frac{1}{2}n}} \exp\left\{-\frac{\Sigma(y_j^2+z_j^2)}{2(1-\rho^2)} + \frac{\rho \Sigma y_j z_j}{1-\rho^2}\right\}, \tag{43}$$

so that the minimal sufficient statistic is $S = \{\Sigma Y_j Z_j, \Sigma(Y_j^2 + Z_j^2)\}$. There does not seem to be an ancillary statistic, i.e. a function of S with a p.d.f. independent of ρ, although $C' = \Sigma(Y_j^2 + Z_j^2)$ is in some reasonable sense approximately ancillary, C' having an expectation $2n$ independent of ρ and variance $4n(1+\rho^2)$ not too strongly dependent on ρ.

If, however, we allow ancillary statistics that are not functions of S, the position is quite different, because both ΣY_j^2 and ΣZ_j^2 separately are ancillary, the corresponding random variables having chi-squared distributions with n degrees of freedom. Clearly there can be no basis for preferring one of ΣY_j^2 and ΣZ_j^2 to the other as an ancillary statistic, so that if the broader definition of ancillarity were adopted, the problem of non-uniqueness would be accentuated.

The definition of ancillary statistics given earlier is restrictive. For instance, in Example 2.25, concerned with random sample size, it is not really crucial that the distribution of sample size should be known. The essential points in that example are that (i) the observed value of sample size by itself should give no information about θ and that (ii) the conditional distribution of the other component given the ancillary statistic depends only on the parameter of interest. The same points arise in connexion with Example 2.27 concerned with random explanatory variables in regression, where the regression parameters γ, β and σ^2 are of primary interest.

To formulate this extended notion of ancillarity, suppose that the unknown parameter θ is partitioned into two parts $\theta = (\psi, \lambda)$, where λ is not of direct interest. We assume that the parameter space is such that any possible value of ψ could arise in conjunction with any possible value of λ, i.e. that $\Omega_\theta = \Omega_\psi \times \Omega_\lambda$, in an obvious notation, the cross denoting Cartesian product. Let S be the minimal sufficient statistic for θ and suppose that $S = (T, C)$, where

(a) the p.d.f. of C depends on λ but not on ψ;

(b) the conditional p.d.f. of T given $C = c$ depends on ψ but not on λ, for all values of c.

Then we call C ancillary for ψ in the extended sense, and T conditionally sufficient for ψ in the presence of the nuisance parameter λ.

With this new definition, we can deal with the situations of Examples 2.25–2.27 when the distributions of sample size, etc. are arbitrary and unknown, or belong to some parametric family, provided that the variation of the ancillary statistic is independent of the parameter of interest in the way just specified.

(ix) Asymptotic sufficiency

In some problems the minimal sufficient statistic may be of dimension n, the number of observations, and yet approximately for large n a statistic of much lower dimension may be "almost sufficient" in

a reasonable sense. This is one aspect of the important matter of finding procedures for complex problems that will have desirable properties asymptotically as $n \to \infty$, and which therefore should have good properties for large but finite n. Chapter 9 develops this topic in detail. Here we give two examples to establish the connexion with sufficiency.

Example 2.31. Maximum likelihood estimates. Suppose that $Y_1, \ldots, Y_n$ are i.i.d. with density $f_{Y_j}(y\,;\theta)$. The asymptotic results of Section 9.2 show that, under certain conditions on $f_Y(y\,;\theta)$, the value $\hat{\theta}$ which maximizes the likelihood is such that for a suitably defined function $i(\theta)$, the likelihood is given over the range of interest by

$$f_Y(y\,;\theta) = f_Y(y\,;\hat{\theta})\left[\frac{\sqrt{n}}{\sqrt{\{2\pi i^{-1}(\theta)\}}}\exp\left\{-\frac{n(\hat{\theta}-\theta)^2}{2i^{-1}(\theta)}\right\} + r_n(y;\theta)\right],$$

where $r_n(y\,;\theta)$ is negligible for large n. Comparing this with the factorization criterion (27), we see that $\hat{\theta}$ satisfies this in the limit and hence can reasonably be called asymptotically sufficient.

Example 2.32. Change-point model (ctd). Suppose $Y_1, \ldots, Y_\gamma$ to be i.i.d. in $N(0, 1)$ and $Y_{\gamma+1}, \ldots, Y_n$ to be i.i.d. in $N(\mu, 1)$ with γ and μ both unknown. Suppose also that there is a restriction on the true value γ, namely $1 \leqslant \gamma \leqslant \gamma_0$ with γ_0 fixed and known. Roughly speaking, for large n we know that there is a change in distribution of the random variables Y_j near the start of the sequence. Then if $\hat{\gamma}$ is the value of γ at which the likelihood is maximized the statistic

$$T_n(\hat{\gamma}) = \sum_{j=\hat{\gamma}+1}^{n} Y_j/(n-\hat{\gamma})$$

is asymptotically sufficient for μ. The heuristic reason is that $T_n(\hat{\gamma})$ and $T_n(\gamma)$ differ by a negligible amount for large n and the latter statistic is sufficient for known γ.

Some explicit details of this problem are given by Hinkley (1972).

2.3 Some general principles of statistical inference

(i) General remarks

In the remainder of this book we develop the theory of a number of types of statistical procedure. One general theme is that the arguments to be used depend both on the type of question of interest

and on the depth to which it is possible to formulate the problem quantitatively. Thus in Chapter 3 we consider situations where only one hypothesis is formulated, whereas in Chapter 11 not only is a full model available for the data, but also there are quantitative specifications of the additional knowledge available and of the consequences of the various possible decisions, one of which is to be chosen in the light of the data. Now it is to be expected on general grounds that once the very much more detailed specification of Chapter 11 is regarded as given, the ideas necessary to develop "optimum" procedures should be relatively straightforward and uncontroversial, whereas when the specification is much weaker there is relatively more need for *ad hoc* arguments and somewhat arbitrary criteria of optimality.

The types of problem that it is worth discussing can be settled only by consideration of applications. We believe that all the levels of specification discussed in the subsequent chapters are useful. Because of this there is no one approach or set of requirements that are universally compelling. The reader may prefer to go straight to the detailed development starting with Chapter 3. On the other hand, there are some general principles that have bearing on the various approaches to be discussed and therefore we now outline these; it is instructive in thinking about particular arguments to consider which of these general principles are obeyed. Some forward reference is inevitable in this and the next section, but has been kept to a minimum.

Throughout, the provisional and approximate character of models has to be borne in mind.

(ii) Sufficiency principle

Suppose that we have a model according to which the observations y correspond to a random variable Y having p.d.f. $f_Y(y\,;\theta)$ and that S is minimal sufficient for θ. Then, according to the sufficiency principle, so long as we accept the adequacy of the model, identical conclusions should be drawn from data y_1 and y_2 with the same value of s.

The argument for this has already been given in Section 2.2 (iii). Once the value of s is known the rest of the data can be regarded as if generated by a fixed random mechanism not depending on, and therefore uninformative about, θ, so long as the assumed model is correct.

A subsidiary but still very important aspect of the sufficiency

principle is that the adequacy of the model can be tested by seeing whether the data y, given $S = s$, are reasonably in accord with the known conditional distribution.

(iii) Conditionality principle

Suppose that C is an ancillary statistic either in the simple sense first introduced in Section 2.2 (viii), or in the second and extended sense where nuisance parameters are present. Then the conditionality principle is that the conclusion about the parameter of interest is to be drawn as if C were fixed at its observed value c. The arguments for this are best seen from Examples 2.26 and 2.27. Suppose that in Example 2.26 it is known that the observations are obtained from $N(\mu, \sigma_1^2)$. How can it affect the interpretation of these data to know that if the experiment were repeated some other variance might obtain? We may think of c as an indicator of which "experiment" was actually performed to produce the data. The following hypothetical example further illustrates the relevance of the conditionality principle.

Example 2.33. Two measuring instruments. A measurement can be taken from one of two measuring instruments $\mathcal{C}_1$ and $\mathcal{C}_2$, with a view to determining whether a physical parameter θ is equal to θ_1 or θ_2. The possible values of the measurement represented by the random variable Y are one and two, such that

$$\text{pr}(Y = 1 \,|\mathcal{C}_1 ; \theta_2) = \text{pr}(Y = 2 \,|\mathcal{C}_1 ; \theta_1) = 1,$$
$$\text{pr}(Y = 1 \,|\mathcal{C}_2 ; \theta_2) = \text{pr}(Y = 2 \,|\mathcal{C}_2 ; \theta_1) = 0.01.$$

The experiment consists of choosing an instrument at random, where $\text{pr}(\text{select } \mathcal{C}_1) = 0.9$ and $\text{pr}(\text{select } \mathcal{C}_2) = 0.1$, and then taking a measurement y. It is known which instrument is used. Suppose now that $y = 1$, and that $\mathcal{C}_2$ was used. Then we calculate that

$$\text{pr}(Y = 1; \theta_1) = 0.099, \quad \text{pr}(Y = 1; \theta_2) = 0.901,$$

which suggests that $\theta = \theta_2$; but the probabilities conditional on $\mathcal{C}_2$ are 0.99 and 0.01, which strongly suggests that $\theta = \theta_1$. In the former, our view is heavily influenced by what might have happened if the more likely instrument $\mathcal{C}_1$ had been used. Thus directly con-

flicting evidence about θ is given if we do not condition on the information "$\mathcal{C}_2$ was used," which is ancillary.

(iv) Weak likelihood principle

With the same information as in (ii), the weak likelihood principle is that two observations with proportional likelihood functions lead to identical conclusions. That is, if y_1 and y_2 are such that for all θ

$$f_Y(y_1;\theta) = h(y_1, y_2) f_Y(y_2 ; \theta)$$

then y_1 and y_2 lead to identical conclusions, so long as we accept the adequacy of the model.

It follows from the construction of Section 2.2(iv) that this is identical with the sufficiency principle.

(v) Strong likelihood principle

Suppose now that two different random systems are contemplated, the first giving observations y corresponding to a vector random variable Y, and the second giving observations z on a vector variable Z, the corresponding p.d.f.'s being $f_Y(y ; \theta)$ and $f_Z(z ; \theta)$ with the same parameter θ and the same parameter space Ω. Then the strong likelihood principle is that if y and z give proportional likelihood functions, the conclusions drawn from y and z should be identical, assuming of course the adequacy of both models. That is, if for all $\theta \in \Omega$

$$f_Y(y ; \theta) = h(y, z) f_Z(z ; \theta), \tag{44}$$

then identical conclusions about θ should be drawn from y and from z.

Examples 2.1–2.3 concerning Bernoulli trials can be used to illustrate this. The log likelihood function corresponding to r successes in n trials is essentially the same whether (a) only the number of successes in a preassigned number of trials is recorded, or (b) only the number of trials necessary to achieve a preassigned number of successes is recorded, or (c) whether the detailed results of individual trials are recorded, with an arbitrary data-dependent "stopping rule". In all cases the log likelihood is, apart from a constant $k(r, n)$,

$$r \log \theta + (n - r) \log(1 - \theta),$$

and if the strong likelihood principle is accepted, then the conclusions

drawn about θ cannot depend on the particular sampling scheme adopted.

These results are very special cases of ones applying whenever we have a "stopping rule" depending in some way on the data currently accumulated but not on further information about the unknown parameter.

Example 2.34. Sequential sampling. Suppose that observations are taken one at a time and that after each observation a decision is taken as to whether to take one more observation. Given $m-1$ observations $y_1, \ldots, y_{m-1}$, there is a probability $p_{m-1}(y_1, \ldots, y_{m-1})$ that one more observation is in fact taken. The conditional p.d.f. of Y_m given $Y_1 = y_1, \ldots, Y_{m-1} = y_{m-1}$ is written in the usual way. Note that this includes very general forms of sequential sampling in which observations may be taken singly or in groups.

Suppose that the data are $(n, y_1, \ldots, y_n)$. Then the likelihood, i.e. the joint probability that observations are taken in the way specified and give the values actually observed, is

$$p_0 f_{Y_1}(y_1; \theta) p_1(y_1) f_{Y_2|Y_1}(y_2|y_1; \theta) \ldots p_{n-1}(y_1, \ldots, y_{n-1})$$

$$f_{Y_n|Y_{n-1},\ldots,Y_1}(y_n|y_{n-1}, \ldots, y_1; \theta)\{1 - p_n(y_1, \ldots, y_n)\}.$$

Thus, so long as the probabilities defining the sampling scheme are known they form a constant factor in the likelihood function and the dependence on the parameters is fixed by the observations actually obtained, in fact by the joint p.d.f. of $Y_1, \ldots, Y_n$. Therefore, if the strong likelihood principle were accepted, the conclusion to be drawn about θ would be the same as if n were fixed. Note, however, that N is not in general an ancillary statistic and that conditioning on its value is not a consequence of the conditionality principle as formulated above.

We noted at the end of the previous subsection that the weak likelihood principle and the sufficiency principle are equivalent. The deduction of the strong likelihood principle from the sufficiency principle plus some form of the conditionality principle has been considered by Birnbaum (1962, 1969, 1970), Barnard, Jenkins and Winsten (1962), Durbin (1970), Savage (1970), Kalbfleisch (1974) and Basu (1973). We shall not go into details, but the following seems the essence of the matter.