

## CHAPTER 1

### STATISTICAL MODELS

#### 1.1 The Theory of Statistical Inference

Often the information collected regarding a phenomenon consists of data that are inherently variable. Thus, it is essential that schemes for interpretation should be available in order to distinguish, in the data, possible underlying regularities from the superficial patterns. The underlying regularities can be conjectured to be effective, the superficial patterns appear clear because of the common psychological distortion that leads to over-rationalization of experience. Both formal and informal tools for interpretation are provided by statistics, which could be depicted as a collection of ideas and methods which aim to describe and evaluate variability and its consequences.

The discipline has developed only rather recently and, encouraged by both internal and external pressures, is still evolving. Among the external pressures are the changing needs and resources of science, technology and society, whilst knowledge of previous progress, the elaboration of hitherto ill-defined analogies and links, the search for clarity, synthesis, speed and efficiency in disseminating important ideas outside of the specialist context are internal pressures.

Naturally, a method that has proved fruitful for one specific problem could, potentially, be of assistance when exploring other problems that are similar either in form or in content. However, analysis of a set of observations always requires a new interpretation and isolated ideas cannot provide a broad enough vision of the data and of the method appropriate for approaching them. The study of statistics, or rather parts of it, can usefully follow structures which tend to unify and prepare for not entirely predictable applications, hence, statistics is not seen as a rigid catalogue of cases and methods.

There are ways of approaching the discipline which aim directly towards

applications, and ways which offer the possibility of an overall view in which details fade and abstraction unites areas that would otherwise be very separate. Focussing on the details of a specific type of observations, and on the appropriate material for understanding them, is typical of applied statistics. Clarifying more general ideas, which are useful for analyzing broad classes of data and for suggesting methods for the analysis of more specific classes of data, is the aim of the theory of statistics.

This book deals in particular with the theory of statistical inference. In this context, the **fundamental assumption** is that the observed data  $y^{obs}$ , often of the form  $y^{obs} = (y_1^{obs}, \dots, y_n^{obs})$ , with  $y_i^{obs}$  an observation on the  $i$ -th observed unit, are the realization of a random vector  $Y$  (or, more generally, of a stochastic process) whose probability distribution is unknown. The way in which this basic abstraction conforms to the observations being studied can vary considerably according to their nature and to the concept of probability that is considered to be the most suitable.

If, as in the frequency view of probability, one wishes to associate probability to a physical meaning, as a way of idealizing the long-run stability of the frequencies observed in a large number of homogeneous repetitions of the generation process, then this requires experimental interpretation of the genesis of the data. This is plausible in many areas of scientific and technological research, indeed, it can sometimes be justified on the basis of randomized allocation of units to treatments, or, in the area of human sciences, whenever a random sampling mechanism is explicitly set up. But, outside of these situations, if one thinks for instance of economic time series, the probability model is essentially an attempt to distinguish systematic and noisy aspects within the variability of the observations.

The context of inference can be summarized by thinking of  $y^{obs}$  as a realization of  $Y \sim p^0(y)$ ,  $y \in \mathcal{Y}$ , where  $p^0(y)$  represents the unknown probability density function (p.d.f.), with respect to a suitable measure, and where  $\mathcal{Y}$  is the sample space. The aim of statistical analysis is to reconstruct  $p^0(y)$  on the basis of both data and suitable assumptions and, possibly, on the grounds of previous information, in order to obtain a concise description of the phenomenon being studied which will permit both interpretation and prediction. Density  $p^0(y)$  or, more accurately, the probability distribution it represents, will be referred to below by the expression **probability model**. Some of the assumptions that facilitate reconstruction of the probability model are usually expressed through a limitation of the possible forms of  $p^0(y)$ , that is, through the specification of a family  $\mathcal{F}$  of probability distributions which are, at least

qualitatively, compatible with  $y^{obs}$ . The family  $\mathcal{F}$  will be called the **statistical model**. If the density function of the probability model which generates the data is an element of  $\mathcal{F}$ , i.e. if  $p^0(\cdot) \in \mathcal{F}$ , the statistical model is said to be **correctly specified**, otherwise the model is said to be **misspecified**.

The probability model has been defined above using the probability density function  $p(y)$ . In some contexts, such a specification would be better expressed in terms of other functions. For instance, if  $Y$  is a univariate random variable (r.v.), we could equivalently describe the distribution of  $Y$  through its distribution function (d.f.)  $F(y) = P(Y \leq y)$ , its moment generating function  $M(t) = E\{\exp(tY)\}$  or its characteristic function  $C(t) = E\{\exp(itY)\}$  (see section 3.2). With a continuous non-negative random variable which describes a lifetime, the **failure rate**  $r(y) = p(y)/\{1 - F(y)\}$  might be more easily interpretable. The quantity  $r(y)dy$  expresses the probability that a failure occurs in the interval  $(y, y + dy)$ , given that it did not occur before  $y$ . If  $Y$  has failure rate  $r(y)$  the probability density of  $Y$  is

$$p(y) = r(y) \exp\left\{-\int_0^y r(t)dt\right\}. \quad (1.1)$$

In the following, when it is necessary to indicate, explicitly, the random variable to which the functions  $p(\cdot)$ ,  $F(\cdot)$ ,  $r(\cdot)$ , etc. pertain, the symbol of the random variable will be written as a subscript, i.e. notations such as  $p_Y(\cdot)$ ,  $F_Y(\cdot)$ ,  $r_Y(\cdot)$  will be used.

## 1.2 Four Paradigms of Inference

From the theoretical point of view, the problem of looking for general principles for statistical inference arises immediately. Principles, that is, which can guide the statistician when seeking suitable techniques for reconstructing  $p^0(y)$ . This search is largely beyond the scope of statistics as it is linked to the more general problem of knowledge and, especially, to the epistemological rules of experimental sciences which are, in the broadest sense, inductive. The literature on this question, although fascinating, has so far proved incapable of offering a satisfactory clarification comparable to that which research on logic has provided for the meaning and the limits of axiomatic structures in mathematics. In other words, the philosophical clarity obtained for deduction has not yet been matched by that for induction. A discussion about foundations of statistical inference is outside the scope of this book. The bibliographic note, section 1.6, gives some references.

However, even in mathematics some scholars, independently of the research in logic, have proposed various conceptions of the foundations of the discipline. Similar investigations have been developed, with even greater urgency, alongside the growth of statistics in the 20th century and have favoured the emergence of some broad structures for interpreting inference. Four general views, or paradigms, for statistical inference are distinguished here, using schematization which is, of course, reductive. The essential differences relate to the interpretation of probability and to the objectives of statistical inference.

The first, which goes back to Bayes and Laplace, is the **personalistic Bayesian paradigm** (or subjectivist paradigm), according to which, interpretation of probability in terms of frequency is only a side-issue, while the fact that it is a description of the subject's state of knowledge is crucial. In the simplest formulation of this view, the subject is required to describe his or her initial state of knowledge in terms of a prior probability distribution on the elements of  $\mathcal{F}$ . Inference is the formalization of how the initial distribution changes in the light of empirical evidence acquired through the data available, according to the one scheme of up-dating that maintains internal consistency, given by Bayes' formula.

Often considered as being out of place in experimental science, subjectivity has been mitigated or disguised by the specification of prior distributions that represent ignorance (Laplace), and hence are somewhat inter-subjective, like the prior distributions deduced from the assumption of equiprobability of the elements of  $\mathcal{F}$ , provided  $\mathcal{F}$  has a finite number of elements. The definition of ignorance or uninformative prior distributions for more complex statistical models was formalized later, from the 1940s on. Some uninformative prior distributions will be considered in sections 2.11.3 and 7.7.2. These later developments identify a **non-personalistic Bayesian paradigm**.

In the early 1920s, in opposition to the Laplace view and, in particular, arguing against the mingling of probability–frequency and probability–opinion, which have an epistemologically different status, Fisher claimed that statistical inference can, indeed must when possible, be *entirely* based on probabilities with experimental interpretation. Here, probability has a dual role: the first is descriptive, that is, modelling variability in a population; the second is epistemological: probability permits quantification of sampling variability of inductions by allowing the variability of samples to be modelled. According to the **Fisherian paradigm**, the variability of inductions should be modelled in accordance with the **principle of repeated sampling**, taking into consideration how the conclusions change with variations in the samples which can be ob-

tained through the hypothetical repetition, under the same conditions, of the experiment which first generated the observations. This hypothetical element refers, in the first place, to the physical non-availability of further samples. It can also be ascribed to the statistical model, in the sense that the behaviour of the inductions is examined for each element of the statistical model. A leading role is played by the concept of likelihood, which is essentially the probability that, within a hypothetical re-running of the experiment, the various competing stochastic mechanisms assign to re-observation of the data that were produced in the actual experiment. A further crucial point in the Fisherian vision is that the probability which describes an event, in order to be relevant, must be considered as conditional on everything that is known. Thus, when judging the probability that a particular seventy year old person will live to the age of seventy-five, not just marginal probability should be used but an attempt should be made to take covariates into account such as sex, state of health, eating habits, family situation and anything else that could prove to be important. Consequently, even when probability is used epistemologically in inference, it must be as relevant as possible to the data  $y^{obs}$ : all the aspects which may be shown by the observations that can be considered as analogous to covariates must be taken into account.

During the 1930s and 1940s, Neyman and Egon Pearson and, later, Wald and Lehmann, contributed to offering a new paradigm for inference that, initially imperceptibly, but later more and more markedly, moved away from the Fisherian view. The starting point was the nucleus of ideas put forward by Fisher and in-depth mathematical study of some of his fundamental concepts, such as likelihood and sufficiency. The emphasis shifted from inference as a summary of data to inferential procedures (hypothesis testing, point and interval estimation) seen as decision problems, in the mathematical form of constrained optimization problems. Further elements which differentiate these new developments were dictated by the need for clarity in mathematical formulations. These required that optimum inference procedures should be identified before the observations were available so as to obey the principle of repeated sampling and leave no shadow of ambiguity about interpretation. This vision will be called the frequency-decision paradigm.

Most of the concepts and methods which will be dealt with here fall within the Fisherian paradigm. Many recent innovations have, directly or indirectly, drawn their inspiration from it. Some such innovations, which have proved as fruitful for theoretical research as they have for applications, are: robust methods, bootstrap, higher-order asymptotic methods and the various con-

cepts of pseudo-likelihood. This does not mean that the ideas of the other three paradigms will be ignored. In particular, classical results of optimal theory of inference (reviewed in section 3.5) will be used; some close relations with the non-personalistic Bayesian paradigm will also be stressed.

### 1.3 Model Specification

It is worth starting from the radically innovative ideas in Fisher (1922a), where the Fisherian paradigm was sketched out for the first time. On the premise that the aim of a statistical analysis is to summarize the data  $y^{obs}$  by means of the reconstruction of  $p^0(y)$ , Fisher divided the problems encountered into three classes:

- **problems of specification**, that are linked to the identification of a statistical model  $\mathcal{F}$  which is appropriate for the observations  $y^{obs}$  being studied; ideally the probability model that generates data is exactly captured by  $\mathcal{F}$  ( $p^0(y) \in \mathcal{F}$ ) or, at least, its most essential aspects are captured;
- **problems of inference**, referred to by Fisher as *problems of estimation*, that is, in general, finding statistical procedures able to locate  $p^0(y)$  within  $\mathcal{F}$  or, with the help of  $\mathcal{F}$ ; if the statistical model  $\mathcal{F}$  is correctly specified, the reconstruction of  $p^0(y)$  will, usually, be all the easier the less mathematically complex  $\mathcal{F}$  is; this class of problems also includes finding procedures which are appropriate for giving indications about model goodness of fit, that is, about the plausibility of the assumption  $p^0(y) \in \mathcal{F}$ ;
- **problems of distribution**, that is the evaluation of how sensitive the reconstruction of  $p^0(y)$  is to the fact that the data used are only a sample; with regard to this, we can presume, in general, that the reconstruction of  $p^0(y)$  will be more effective the lower the extension of  $\mathcal{F}$  is.

The three classes of problems (specification, inference and distribution) are closely interlinked. In applications, use of a model  $\mathcal{F}$  for which the available inference methods are either difficult to use or not particularly effective is discouraged. Vice versa, any substantial progress related to the problems of the second and third classes, facilitates the solution of specification problems and broadens the reserve of statistical models that can, reasonably, be drawn upon. These interactions are complementary, in the sense that interest in the

application of statistical models which pose complex problems of inference is a stimulus to further research.

In applications, a statistical model is usually considered as an approximation, in the sense that no one expects it to capture the probability model accurately, rather, the idea is that the model can be considered adequate for the aims of the research. In this sense, the three classes of problems should not be understood as corresponding to successive phases in data analysis, rather, they should be understood as logical moments along a necessarily iterative path. This same specification phase is usually, to a greater or lesser extent, guided by the data through the feedback effects of the inference phases: as, for example, in the analysis of residuals in a linear regression model.

Although model specification is very important, and usually reflects more on the conclusions than does the inference paradigm adopted, the theory of statistical inference, traditionally, lacks explicit indications about specification problems. Fisher, who probably felt that the unique aspects of any single type of data were pre-eminent, assigned these problems to applied statistics. This vision seems to be an over-simplification because it introduces a split between theory and application. Theory puts some particular models in the spotlight and, even though it is true that the specification phase is hard to formalize, some guidelines, based on common sense, should be offered.

A starting point could be that of setting limitations which are based upon the nature of the data being examined. Different models are suitable for dealing with qualitative (nominal, ordinal) or quantitative (discrete, continuous) variables, and with functions, images, etc.. In addition, the observed variables could be subdivided into subsets, for example, into response and explanatory variables. In order to be specified reasonably accurately, the model must respect both the support and the role of variables.

A second point is that the information about the observation scheme is also important: for example, idealizations such as random sampling, randomization, censoring or other models for missing data, sequential sampling, time and space dependence.

Furthermore, attention should be focussed on aspects of the data that the model must be able to catch, for example, the centre of distribution, unimodality or bimodality, dependence on explanatory variables etc., and complementary aspects, such as dispersion, asymmetry, heteroschedasticity, etc., which should be taken into account in order to perfect the analysis. The statistical model must be able to succinctly describe the aspects that are of primary interest and must also be sufficiently flexible to allow a realistic description of

additional aspects.

### 1.3.1 Levels of Specification

Depending on the information available it may be deemed appropriate to extend the statistical model  $\mathcal{F}$  to a greater or a lesser degree. In increasing order of extension, hence in decreasing order of presumed information, the following three levels of specification can be outlined.

- **Parametric specification.** Previous knowledge and conjecture produce a rather restricted class  $\mathcal{F}$ ; the elements that make it up can be indexed by a finite number  $p$  of real parameters, that is

$$\mathcal{F} = \{p(y; \theta), \theta \in \Theta \subseteq \mathbb{R}^p\}.$$

If the model is correctly specified, we have  $p^0(y) = p(y; \theta_0)$  for a value  $\theta_0 \in \Theta$ , called the **true parameter value**.

- **Semiparametric specification.** The elements of  $\mathcal{F}$  can be identified through both a parametric and a nonparametric component, that is,  $\mathcal{F} = \{p(y) = p(y; \theta), \theta \in \Theta\}$  where  $\theta = (\tau, h(\cdot))$ , with  $\tau \in \mathbb{T} \subseteq \mathbb{R}^k$  whereas the set of possible specifications of the function  $h(\cdot)$  cannot be indexed by a finite number of real parameters. In this specification,  $\tau$  usually represents the aspects of primary interest of the distribution, while high flexibility is sought in the description of the aspects of secondary concern. Think, for example, of the class of continuous symmetrical distributions on  $\mathbb{R}$ , which are partially parameterized by their centre of symmetry, that is, have density of the form  $p(y; \mu) = p_0(y - \mu)$ , with  $\mu \in \mathbb{R}$  and  $p_0(\cdot)$  a probability density symmetric around the origin. Furthermore, think of the usual linear regression model of  $Y$  on  $X$ ,  $E(Y_i) = \alpha + \beta x_i$ , where the  $Y_i$ ,  $i = 1, \dots, n$ , are independent with a common variance  $\sigma^2$ , the  $x_i$  are  $n$  known constants,  $\tau = (\alpha, \beta, \sigma^2)$ , and the distribution of  $Y$  is not further specified. One last, rather important, example is the **proportional hazards model** (Cox, 1972), which is suitable for expressing the dependence of the lifetime  $Y_i$  of the  $i$ -th unit on a  $k$ -dimensional explanatory variable  $\mathbf{x}_i$ . This model is specified through the failure rate function as

$$r_{Y_i}(y_i) = r_0(y_i) \exp\{\beta \cdot \mathbf{x}_i\}, \quad (1.2)$$

where  $r_0(\cdot)$  is an unknown **baseline hazard function**,  $\beta = (\beta_1, \dots, \beta_k)$  denotes a vector of regression coefficients, which are also unknown, and  $\beta \cdot \mathbf{x}_i$  denotes the scalar product of the two vectors.

- **Nonparametric specification.** The model  $\mathcal{F}$  is a restriction of the family of all the probability distributions defined on a support which is suitable for the data under analysis. It is defined by means of global simplifying assumptions which do not expressly identify a finite number of parameters that are primary subject of inference. For example if the observations are  $y = (y_1, \dots, y_n)$ , a possible nonparametric model is given by the restriction of the distributions on  $\mathbb{R}^n$  to the subfamily  $\mathcal{F}$  which is made up of distributions with independent and identically distributed components (random sampling of size  $n$  from a random variable with an entirely unknown distribution).

Because of the, usually unavoidable, element of approximation that is inherent in a statistical model, it is often a guide rather than a prescription. The choice of the level of specification depends, among other factors, both upon how much information one can reasonably expect to extract from the data and on the aim for which the model is used. When there is only a small amount of data a rather parsimonious parametric model can be useful. If the model is purely empirical, a black box that directs action, then its mathematical manageability is of prime importance. If, on the other hand, it must explain the subject being studied and represent an element for understanding, then a parametric model is usually required. This model should be carefully deduced from simple assumptions elicited through a fundamental examination of the mechanism that generates the data. Usually, in this type of model, each parameter represents one specific aspect endowed with a physical meaning. One strategy that is always useful when building a complex model is that of combining elementary blocks. For example, it is worth separating any consideration of the deterministic part from that of the stochastic part. Usually, the former can be more closely linked to the essential knowledge or to the aims of the analysis and then modelled parametrically, while, in the latter, it is better to avoid over-restrictive assumptions. Just think of the description of the effect of covariates carried out through regression analysis, that is, through modelling the connection between the parameters of the distribution of the response variable with the values of the covariates.

### *1.3.2 Notes on the Specification of a Parametric Model*

The problem of specification is always present but is enhanced in the parametric case. Here, a direct comparison is required between the knowledge

available regarding the mechanism that has generated the data and the probabilistic genesis, exact or asymptotic, of the various families of distributions. The binomial distribution describes the random variability of the number of successes in a given number of trials *if* the trials are independent *and if* the probability of success in each trial is constant. These assumptions give a suitable statistical model if they are met to a reasonable degree of approximation. Such a statistical model could still be useful even if the approximation is not accurate, so long as the usual inferential procedures are suitably modified, e.g. so that they can take lack of independence or overdispersion produced by the heterogeneity of trials into account. When fitting a continuous distribution to time to failure, the exponential distribution is useful to describe populations that do not age, in the sense that if an item works at a given time, the distribution of its additional lifetime is the same as that of new items (see for example, Azzalini, 1996, section A.2.4). If essential knowledge suggests ageing for the phenomenon being studied, the data should be modelled through specification of the failure rate.

More generally, characterization results that express, in a simple form, which aspects will emerge from the data under a particular model can also be helpful. A characterization of a family of probability distribution  $\mathcal{F}$  states a necessary and sufficient condition for the density  $p(y)$  to belong to  $\mathcal{F}$ . For example, a constant failure rate characterizes the exponential distribution in the class of continuous distributions. As a further example, the assumption that  $Y_1, \dots, Y_n$  are independent and identically distributed (i.i.d.) and follow a normal distribution with mean 0 and variance  $\sigma^2$ ,  $N(0, \sigma^2)$ , is equivalent to the assumption that their joint density  $p_y(y)$  can be factorized according to stochastic independence and is spherically symmetrical with respect to the origin of  $\mathbb{R}^n$  (see e.g. Lehmann, 1990, Example 2.1).

Some parametric models can be justified on the basis of asymptotic considerations. Besides the well known case of the normal distribution, which is tied to the central limit theorem, **extreme value distributions** are also defined by limiting arguments. If the maximum of  $n$  independent observations from a univariate distribution, suitably standardized, has a limiting distribution as  $n$  tends to infinity, then this limiting distribution can only belong to one of three families. These distributions are useful when describing phenomena that could be interpreted in terms of a large value of either a detectable or a hidden variable (maximum annual rainfall, failure time, etc.). The asymptotic results relating to stable distributions (see section 3.2.6) have also attracted some interest in statistics. Stable distributions arise from generalizations of the central

limit theorem to the infinite variance case. In many applications a model that permits infinite variance in the observable variables is unsuitable. Exceptions can be found in models for economic and financial data, see Du Mouchel (1983). Basic convergence results for sums and extremes are collected in Appendices A and B, respectively.

The connection between the genesis of a parametric model or characterization results and the data may be weak or unclear. Asymptotic arguments require an evaluation of the adequacy of the approximation given. The assumptions of independence and identical distribution, which may have been made hastily, should be critically examined. Furthermore, particularly in the specification of complex models, it may happen that some parts of the model can be expressed with reasonable confidence, while others require closer examination in the light of the information provided by the data. All these elements make it clear that the specification of a model is usually the product of an iterative process. The choice between competing models can be made through informal and formal tools, such as plots, analysis of residuals, selection procedures and tests of goodness-of-fit.

The rest of this book will mainly deal with problems of inference and of distribution and will only offer the reader indirect help with specification problems. It will, possibly, broaden the range of models which could be considered familiar, both with their definition and the mastery of their salient inferential properties. Even though complex models will only rarely be expressly mentioned, in-depth understanding of the blocks from which they are built will facilitate their definition and study in the specific applications of statistics.

## 1.4 Parametric Statistical Models and Likelihood

Assume that the family  $\mathcal{F}$  has been specified as a class of probability models compatible with  $y^{obs}$ . Of course, below, this assumption is understood as being provisional, it is a working hypothesis that is liable to be reformulated in the light of any new information which becomes available in later stages of statistical analysis.

### 1.4.1 General Formulation of a Statistical Model

The statistical model  $\mathcal{F}$  is usually described as a family of density functions. A mathematically precise formulation requires notions of measure theory. Measure theory is not essential to most of the topics treated in this book; however,

the reader is assumed to be familiar with the basic definitions and results, at the level, for instance, of section 1.2 in Lehmann (1983).

In general terms, the class  $\mathcal{F}$  is specified once the triple

$$(\mathcal{Y}, P_\theta, \Theta)$$

has been assigned, where  $\mathcal{Y}$  denotes the sample space,  $P_\theta$  a probability measure on a  $\sigma$ -algebra  $\mathcal{B}$  of subsets of  $\mathcal{Y}$ , called events. The distribution  $P_\theta$  is indexed by the parameter  $\theta$ , which takes values in the parameter space  $\Theta$ . If  $\Theta$  is, abstractly, meant as a set of indices, then both nonparametric, semiparametric and parametric specifications fall within this formulation. It is assumed that the **identifiability condition** is met,  $P_\theta \neq P_{\theta'}$  if  $\theta \neq \theta'$ . This means that there is at least one event  $B \in \mathcal{B}$  such that  $P_\theta(B) \neq P_{\theta'}(B)$ .

Even though some of the ideas presented below could be interpreted as referring to this more general specification, attention will be concentrated on classes  $\mathcal{F}$  of parametric models for which the following further conditions are satisfied.

- There exists a  $\sigma$ -finite measure  $\mu^*$  on  $\mathcal{B}$  such that all the probability measures  $P_\theta$ ,  $\theta \in \Theta$ , are **absolutely continuous** with respect to  $\mu^*$ , that is, whichever  $\theta \in \Theta$ ,  $P_\theta(B) = 0$  for every event  $B$  for which  $\mu^*(B) = 0$ . The family  $\mathcal{F}$  is then called **dominated**. Then, by the Radon–Nikodym theorem, the density function of  $P_\theta$  with respect to the dominating measure  $\mu^*$ ,  $p_Y(y; \theta) = dP_\theta/d\mu^*$ , is defined. The probability of  $B \in \mathcal{B}$  may be expressed as  $P_\theta(Y \in B) = \int_B p_Y(y; \theta) d\mu^*$ . In most applications  $\mathcal{Y}$  is a subspace of a Euclidean space and  $\mu^*$  is the Lebesgue measure (absolutely continuous case) or a counting measure (discrete case), so that  $P_\theta(Y \in B) = \int_B p_Y(y; \theta) dy$  or  $P_\theta(Y \in B) = \sum_{y \in B} p_Y(y; \theta)$ , respectively.
- The parameter space  $\Theta$  is a subset, possibly the entire space, of the  $p$ -dimensional Euclidean space  $\mathbb{R}^p$ . In the simplest setting one assumes that  $\Theta$  is an open non-empty subset of  $\mathbb{R}^p$ ; on one hand, this makes the treatment of aspects of differential calculus for smooth functions defined on  $\Theta$  easier, on the other hand, it gives  $p$  the meaning of an effective geometric dimension of  $\Theta$ .

Thus, a dominated parametric statistical model  $\mathcal{F}$  can be specified in the form

$$(\mathcal{Y}, p_Y(y; \theta), \Theta), \tag{1.3}$$

where  $p_Y(y; \theta)$  is a density with respect to  $\mu^*$  and  $\Theta \subseteq \mathbb{R}^p$ .

Below, especially when theoretical considerations are presented, the components of  $\theta$  are indicated by  $\theta^1, \dots, \theta^p$ . It should be noted that with this notation,  $\theta^2$  is the second component of  $\theta$  and not  $\theta$  squared. Furthermore, the symbols  $\theta^r, \theta^s$ , etc.,  $r, s, \dots = 1, \dots, p$ , are used to indicate generic components of  $\theta$ . The notational convention of using upper indices for the components of  $\theta$  may be new to the reader. The advantages of using this notation will become clearer later.

One fact which should be underlined is that, by attributing to  $\theta$  the mere function of indicator of elements in  $\mathcal{F}$ , it makes no difference if it is replaced by a one-to-one smooth function (infinitely differentiable with infinitely differentiable inverse), considering a reparameterization for  $\mathcal{F}$ . In applications, it might be more suitable to choose a parameterization in which every component of the parameter describes an easy to interpret characteristic of distribution. On the other hand, inference could be simplified by the choice of a different parameterization, without changing the essence of the problem which remains that of locating  $p^0(y)$  within  $\mathcal{F}$ .

If the data consist of  $n$  observations  $y = (y_1, \dots, y_n)$ , both the sample space  $\mathcal{Y}$  and the density  $p_Y(\cdot)$  depend on  $n$ . As regards the parameter space, we will mainly deal with situations where the dimension of  $\Theta$  does not depend on  $n$ , thus excluding, at least initially, having to deal with models with incidental parameters (see section 4.1). If the  $n$  observations are independent, the sample space  $\mathcal{Y}_n$  is the Cartesian product of the sample spaces of the individual observations, the density is  $p_Y(y; \theta) = \prod_{i=1}^n p_{Y_i}(y_i; \theta)$ , with  $p_{Y_i}(y_i; \theta)$  density of the  $i$ -th observation. If the  $n$  observations are dependent, it is usually convenient to write their joint density in product form, as

$$p_Y(y_1, \dots, y_n; \theta) = p_{Y_1}(y_1; \theta) p_{Y_2|Y_1=y_1}(y_2; y_1, \theta) \\ \cdots p_{Y_n|Y_1=y_1, \dots, Y_{n-1}=y_{n-1}}(y_n; y_1, \dots, y_{n-1}, \theta),$$

where  $p_{X|Z=z}(x; z)$  denotes here, and in the following, the conditional density of  $X$  given  $Z = z$ .

#### 1.4.2 Likelihood and Related Quantities

The concept of likelihood is crucial for the Fisherian view of parametric statistical inference.

**Definition 1.1** Let  $\mathcal{F}$  be a parametric statistical model for data  $y$  specified in the form (1.3). The likelihood function is

$$L = L(\theta) = L(\theta; y) = c(y) p_Y(y; \theta), \quad (1.4)$$

where  $\theta \in \Theta$  and  $c(y) > 0$  is an arbitrary constant of proportionality.

The function  $L(\theta; y)$  gives the natural information summary on  $\theta$ , based on the parametric statistical model (1.3) and on the observed data  $y$ . The arguments for this statement will be discussed in Chapter 2. Even though likelihood was introduced into a strictly parametric context, recent developments have demonstrated that the concept of likelihood is productive in the context of semi-parametric and nonparametric models too (see sections 4.8-4.10). The term *likelihood* was first introduced in statistics by Fisher (1921). In Fisher (1922a) the following concise definition was given.

*Likelihood.*—*The likelihood that any parameter (or set of parameters) should have any assigned value (or set of values) is proportional to the probability that if this were so, the totality of observations should be that observed.*

The definition of the likelihood function up to constant factors can be justified in various ways. At an intuitive level, the fact that likelihood serves as an indicator of coherence between probability models, which correspond to the possible values of the parameter, and observations, means that only relative comparisons are possible, where the constant is irrelevant. Moreover, the definition of  $L(\theta)$  is independent of the dominating measure  $\mu^*$ . This has two aspects. Firstly, as in the discrete case where  $L(\theta; y)$  is proportional to the probability of re-observing  $y$  in a hypothetical repetition of the experiment, hence this interpretation is valid in the continuous case, according to a clear limiting procedure. Secondly, the likelihood function does not change under one-to-one transformations of the data; in the continuous case, the Jacobian determinant does not depend on the parameter and is incorporated in the constant of proportionality. For a further reason, see Example 2.4. One important consequence of the presence of  $c(y)$  in the definition (1.4) is that the likelihood function is independent of the sampling rule provided that this rule depends only on the data and not on  $\theta$  (see Problem 2.2).

When dealing with the likelihood function, it may in many cases prove more convenient to consider the **log-likelihood function**

$$l = l(\theta) = l(\theta; y) = \log L, \quad (1.5)$$

where  $\log(\cdot)$  denotes the natural logarithm and  $l(\theta) = -\infty$  if  $L(\theta) = 0$ . Since  $L(\theta)$  is defined up to a multiplicative constant, the log-likelihood is, in its turn, determined up to an additive constant, which only depends on  $y$ . In the case of independent observations

$$l(\theta) = \sum_{i=1}^n \log p_{Y_i}(y_i; \theta). \quad (1.6)$$

Adopting the repeated sampling principle as a criterion of evaluating inferential procedures requires study of the probability distribution of the random variable  $l(\theta) = l(\theta; Y)$ , or of related quantities, for  $\theta$  fixed and as  $y$  varies in the sample space  $\mathcal{Y}$  according to a density  $p_Y(y; \tilde{\theta})$  in  $\mathcal{F}$ , where  $\tilde{\theta} \in \Theta$  is a parameter value not necessarily equal to  $\theta$ . Generally, we speak of a **null distribution** if  $\tilde{\theta} = \theta$  and of a **non-null distribution** if  $\tilde{\theta} \neq \theta$ . Analogously, we refer to **null moments**, that is evaluated with respect to a null distribution, and to **non-null moments**, evaluated with respect to a non-null distribution. Symbols such as  $E_{\theta}(\cdot)$ ,  $\text{Var}_{\theta}(\cdot)$  are used to indicate expectation, variance (or covariance matrix), etc., calculated with reference to  $p_Y(y; \theta)$ . In some cases, it must be explicitly indicated which distribution has been used to calculate expectation, variance, etc.; in these cases,  $E_{Y; \theta}(\cdot)$ ,  $\text{Var}_{Y; \theta}(\cdot)$  and other analogous notations are used. It should be noted that while the likelihood function is not sensitive to the sampling rule, the distributions and the moments introduced above depend on hypothetical repetitions of the experiment.

Jensen's inequality offers a basic result for the expectation of log-likelihood which will be referred to as the **Wald inequality**:

$$E_{\theta}(l(\theta)) > E_{\theta}(l(\tilde{\theta})), \quad \tilde{\theta} \neq \theta. \quad (1.7)$$

According to (1.7), the null expectation of log-likelihood is always greater than the non-null expectation.

In the following, it will be assumed that log-likelihood is a sufficiently smooth function of  $\theta$ , that is, that it has partial derivatives with respect to the components of  $\theta$  up to the required order. Furthermore, it is assumed that all the null moments of these derivatives are finite. If non-null moments are also referred to, then their existence is implicit. The partial derivatives of the log-likelihood function are indicated by

$$l_r = l_r(\theta; y) = \frac{\partial}{\partial \theta^r} l(\theta),$$

$$l_{rs} = l_{rs}(\theta; y) = \frac{\partial^2}{\partial \theta^r \partial \theta^s} l(\theta),$$

$$l_{rst} = l_{rst}(\theta; y) = \frac{\partial^3}{\partial \theta^r \partial \theta^s \partial \theta^t} l(\theta),$$

etc.. Such derivatives, and more generally the quantities obtained from the likelihood function will be called **likelihood quantities**. The derivatives of  $l(\theta)$  up to the second order, play a crucial role in inference. Derivatives up to the fourth order and their joint moments are important for refining the asymptotic theory of inference (see Chapters 9 and 11). Likelihood quantities that depend on the observed data are called **observed likelihood quantities** and functions of their moments are called **expected likelihood quantities**.

**Definition 1.2** The **score function**  $l_*$  is the vector of the partial derivatives of  $l(\theta)$  with respect to  $\theta$ , i.e.  $l_* = l_*(\theta; y) = (l_1, \dots, l_p)$ .

The notation  $l_* = [l_r]$  will also be used, where  $[a_r]$  denotes a vector with generic element  $a_r$  and the range of the index  $r$  is understood.

Assuming that the conditions which permit differentiation and integration to be interchanged are satisfied, from the identity

$$\int_{\mathcal{Y}} p_Y(y; \theta) d\mu^* = 1$$

it follows that

$$E_\theta(l_*) = E_\theta(l_*(\theta; Y)) = 0, \quad (1.8)$$

that is, the null first moment of the score is zero. This is what one would expect from (1.7).

**Definition 1.3** The **observed information matrix**,  $j(\theta)$ , is

$$j = j(\theta) = \begin{pmatrix} -l_{11} & \cdots & -l_{1p} \\ \vdots & \ddots & \vdots \\ -l_{p1} & \cdots & -l_{pp} \end{pmatrix},$$

or, in compact notation,  $j = [-l_{rs}]$ ; here and below the matrix with elements  $a_{r,s}$  is indicated by  $[a_{r,s}]$ . Furthermore,

$$i = i(\theta) = E_\theta\{j(\theta)\} = [E_\theta\{-l_{rs}(\theta; Y)\}]$$

denotes the **expected information** or **Fisher information matrix**.

It is assumed here that the conditions which ensure the validity of the information identity

$$E_{\theta}\{-l_{rs}(\theta)\} = E_{\theta}\{l_r(\theta)l_s(\theta)\}, \quad (1.9)$$

$r, s = 1, \dots, p$ , are satisfied (see e.g. Azzalini, 1996, section 3.2.4). In other words, the expected information matrix is the null second moment of the score and, as such, is a non-negative definite matrix. The elements of the inverse of a matrix  $[a_{rs}]$  will be denoted by upper indices, that is,  $[a_{rs}]^{-1} = [a^{rs}]$ ; in particular  $j^{-1} = [j^{rs}]$  and  $i^{-1} = [i^{rs}]$ , when these inverses exist.

The sequence with three indices of third order partial derivatives of log-likelihood,  $l_{rst}$ , can be considered collectively as an aggregate of objects endowed with the relevant order structure, called *array*, indicated by  $[l_{rst}]$ . Analogous notations may be used for higher-order derivatives. Generally, in an array, the indices can appear both as upper and lower indices; a specific meaning is attributed to the position of an index in section 9.5. For the moment, attention should be drawn to the fact that an expression such as  $l_{rst}$  denotes a generic element of  $[l_{rst}]$ , that is, a generic function which appears in such an array; the range of the indices ( $r, s, \dots = 1, \dots, p$ ) is usually understood.

The likelihood function highlights some probability models included in the parametric statistical model  $\mathcal{F}$  as being particularly suitable for interpreting the variability observed.

**Definition 1.4** A value of  $\theta$  that maximizes  $L(\theta; y)$  over  $\Theta$ , that is a value  $\hat{\theta}$  such that  $L(\hat{\theta}) = \sup_{\theta \in \Theta} L(\theta)$ , is called **maximum likelihood estimate (m.l.e.)** of  $\theta$ .

Of all the elements of  $\mathcal{F}$ ,  $\hat{\theta}$  denotes that (or those) for which data  $y$  offer the maximum empirical support. The notations  $\hat{l} = l(\hat{\theta})$ ,  $\hat{j} = j(\hat{\theta})$ ,  $\hat{i} = i(\hat{\theta})$ , etc., are used to denote likelihood quantities evaluated at  $\hat{\theta}$ . By slightly adjusting the terminology, if  $\theta = (\tau, \zeta)$  and  $\hat{\theta} = (\hat{\tau}, \hat{\zeta})$ , we say that  $\hat{\tau}$  is the maximum likelihood estimate of  $\tau$ .

Maximizing  $l(\theta)$  instead of  $L(\theta)$  is often an advantage in applications, and, whenever  $l(\theta)$  can be differentiated, it is worth seeking the maximum likelihood estimate from among the solutions of the **likelihood equation**

$$l_{*}(\theta) = 0. \quad (1.10)$$

In the following, unless otherwise stated, it will be assumed that the maximum likelihood estimate is unique and that it is the unique solution to (1.10).

Sufficient conditions for specific classes of parametric models will be given in Chapters 5, 6 and 7. Equation (1.10) is an **estimating equation** naturally associated with the parametric model  $\mathcal{F}$ .

Under a parametric statistical model  $\mathcal{F}$  with data  $y^{obs}$ , inference about  $\theta$  can be carried out in qualitative terms on the basis of the behaviour of the likelihood function. Intuitively, the true parameter value  $\theta_0$  is located close to the value  $\hat{\theta}$  which has maximum empirical support in terms of likelihood. Furthermore, the rapidity of the decrease of log-likelihood around its maximum which is measured, for example, by  $\hat{j}$ , is a source of information about the variations in the empirical support for  $\theta$  values which are close to  $\hat{\theta}$ , in the sense that any drastic variations are a sign of efficacious localization. Lastly, further aspects such as the asymmetry of  $l(\theta)$  around the global maximum, the presence of local maxima, slow decrease on one tail, and so on, give an articulated vision of the information gleaned from the data. The practise of examining the log-likelihood function is to be recommended and it is facilitated by the graphic options of many popular statistics packages. For an approach to inference which is based entirely on the likelihood function, see Edwards (1972).

### 1.4.3 Reparameterizations

The likelihood and the log-likelihood functions do not depend on the parameterization chosen for  $\mathcal{F}$ . In fact, let  $\psi = \psi(\theta)$  be a one-to-one smooth function from  $\Theta \subseteq \mathbb{R}^p$  to  $\Psi \subseteq \mathbb{R}^p$ , infinitely differentiable together with its inverse. Then  $\psi$  defines an alternative parameterization of the model. Since  $\theta$  and  $\psi(\theta)$  identify the same element of  $\mathcal{F}$ , we can write

$$L^\Psi(\psi) = L^\Theta(\theta(\psi)) \quad (1.11)$$

and

$$l^\Psi(\psi) = l^\Theta(\theta(\psi)), \quad (1.12)$$

where likelihood and log-likelihood in the parameterization  $\psi$  are denoted by  $L^\Psi(\cdot)$  and  $l^\Psi(\cdot)$  and the same functions in the original parameterization  $\theta$  are denoted by  $L^\Theta(\cdot)$  and  $l^\Theta(\cdot)$ ; the inverse function of  $\psi(\theta)$  is denoted by  $\theta(\psi)$ .

The transformation law (1.11) excludes any interpretation of likelihood, multiplied by the suitable normalizing constant, as a probability distribution on  $\Theta$ . In this case, for continuous parameters,  $L^\Theta(\cdot)$  and  $L^\Psi(\cdot)$  should differ through the Jacobian determinant of the transformation  $\psi(\theta)$ .

On the basis of relations (1.11) and (1.12), likelihood may be considered as an intrinsic function, i.e. one not dependent on the coordinate system expressed by the parameterization, defined on the collection of probability models  $\mathcal{F}$ . Indeed, it is possible to define the likelihood function as  $L(p(\cdot))$ ,  $p(\cdot) \in \mathcal{F}$ . When a parameterization is assigned as a function  $\theta : \mathcal{F} \rightarrow \mathbb{R}^p$ , then the expression  $L(\theta)$  should be understood as the abbreviated form of  $L(\theta(p(\cdot)))$ . Analogously, the expression  $l(\theta)$  is the abbreviated form of  $l(\theta(p(\cdot)))$ . Even though the notations  $L(\theta)$  and  $l(\theta)$  will, henceforth, be adopted, it should be remembered that every function of  $L(\theta)$  and  $l(\theta)$  can be defined directly on  $\mathcal{F}$  and only for convenience is expressed in terms of a given parameterization. One important example is the maximum likelihood estimate:  $\hat{\theta}$  indicates the element of  $\mathcal{F}$  for which the data offer maximum support. If such an element is indicated by  $\hat{p}(\cdot)$  it will be  $\hat{\theta} = \theta(\hat{p}(\cdot))$ . From this follows the important property of equivariance under reparameterization of the maximum likelihood estimate. If  $\psi$  is an alternative parameterization of  $\mathcal{F}$ , we will have  $\hat{\psi} = \psi(\hat{p}(\cdot)) = \psi(\hat{\theta})$  and  $\hat{\theta} = \theta(\hat{\psi})$ .

On the other hand, the log-likelihood derivatives  $l_r, l_{rs}, \dots$ , and their expectations depend on the parameterization chosen for  $\mathcal{F}$  and they transform according to regular patterns under a reparameterization. Let us denote by  $\psi^a, \psi^b, \dots$  ( $a, b = 1, \dots, p$ ), the generic components of  $\psi$ , whereas  $\theta^r, \theta^s, \dots$ , denote the generic components of  $\theta$ . According to the differentiation rule for composite functions, the relation between the scores in the two parameterizations can be expressed as

$$\bar{l}_a = \sum_{r=1}^p l_r \theta_a^r, \quad (1.13)$$

where

$$\bar{l}_a = \frac{\partial}{\partial \psi^a} l^\Psi(\psi), \quad (1.14)$$

$l_r = l_r(\theta(\psi))$  and  $\theta_a^r = (\partial/\partial \psi^a)\theta^r(\psi)$ . Let

$$\bar{l}_{ab} = \frac{\partial^2}{\partial \psi^a \partial \psi^b} l^\Psi(\psi), \quad (1.15)$$

$$\bar{l}_{abc} = \frac{\partial^3}{\partial \psi^a \partial \psi^b \partial \psi^c} l^\Psi(\psi), \quad (1.16)$$

etc., denote partial log-likelihood derivatives in the parameterization  $\psi$ . By reapplying the differentiation rule for composite functions the following further

relationships are obtained

$$\bar{l}_{ab} = \sum_{r,s=1}^p l_{rs} \theta_a^r \theta_b^s + \sum_{r=1}^p l_r \theta_{ab}^r, \quad (1.17)$$

$$\bar{l}_{abc} = \sum_{r,s,t=1}^p l_{rst} \theta_a^r \theta_b^s \theta_c^t + \sum_{r,s=1}^p l_{rs} (\theta_{ab}^r \theta_c^s + \theta_{ac}^r \theta_b^s + \theta_{bc}^r \theta_a^s) + \sum_{r=1}^p l_r \theta_{abc}^r, \quad (1.18)$$

etc., where  $l_{rs} = l_{rs}(\theta(\psi))$ ,  $l_{rst} = l_{rst}(\theta(\psi))$ ,  $\theta_{ab}^r = (\partial^2/\partial\psi^a \partial\psi^b)\theta^r(\psi)$  and  $\theta_{abc}^r = (\partial^3/\partial\psi^a \partial\psi^b \partial\psi^c)\theta^r(\psi)$ .

Information in the parameterization  $\theta$  and information in the parameterization  $\psi$  are linked by the identities

$$\bar{j}_{ab} = \sum_{r,s=1}^p j_{rs} \theta_a^r \theta_b^s - \sum_{r=1}^p l_r \theta_{ab}^r, \quad (1.19)$$

$$\bar{i}_{ab} = \sum_{r,s=1}^p i_{rs} \theta_a^r \theta_b^s. \quad (1.20)$$

Perhaps the reader would be more familiar with the two expressions (1.13) and (1.20) in matrix notation, in the form

$$I_*^\Psi(\psi) = [\theta_a^r]^\text{T} I_*^\Theta(\theta(\psi))$$

and

$$i_*^\Psi(\psi) = [\theta_a^r]^\text{T} i_*^\Theta(\theta(\psi)) [\theta_a^r], \quad (1.21)$$

where  $\Theta$  and  $\Psi$  denote the reference parameterization,  $[\theta_a^r]$  denotes the  $p \times p$  matrix with  $\theta_a^r$  as element of position  $(r, a)$  and the symbol T is used to denote transposition. As will be discussed later in sections 9.1 and 9.2, matrix notation is not really suitable for expressing, in compact form, the rules of transformation of quantities which are linked to derivatives of order higher than two, as for example in (1.18); therefore, it will be convenient to use a different notation.

### Example 1.1 Poisson distribution

Let  $y_1, \dots, y_n$  be independent and identically distributed observations drawn from a Poisson distribution with mean  $\theta$ . The log-likelihood function is

$$l(\theta) = \log(\theta) \sum_{i=1}^n y_i - n\theta,$$

$\theta > 0$ ,  $y_i = 0, 1, \dots$ . The score function and the expected and observed information are

$$\begin{aligned} l_*^\ominus(\theta) &= l_1 = \frac{1}{\theta} \sum_{i=1}^n y_i - n, \\ j^\ominus(\theta) &= \frac{1}{\theta^2} \sum_{i=1}^n y_i, \\ i^\ominus(\theta) &= \frac{n}{\theta}. \end{aligned}$$

Consider the new parameterization  $\psi = \psi(\theta) = e^{-\theta} = p_{y_i}(0; \theta)$ , with inverse  $\theta = \theta(\psi) = -\log \psi$ . Instead of calculating the quantities  $l_*^\Psi(\psi)$ ,  $j^\Psi(\psi)$  and  $i^\Psi(\psi)$  starting from the new parameterization, the relationships (1.13), (1.19) and (1.20) can be used, by taking into account that  $(\partial/\partial\psi)\theta(\psi) = \theta_1^1 = -1/\psi$  and  $(\partial^2/\partial\psi^2)\theta(\psi) = \theta_{11}^1 = 1/\psi^2$ , thus directly obtaining,

$$\begin{aligned} l_*^\Psi(\psi) &= \left( \frac{\sum y_i}{-\log \psi} - n \right) \left( -\frac{1}{\psi} \right) = \frac{1}{\psi} \left( \frac{\sum y_i}{\log \psi} + n \right), \\ j^\Psi(\psi) &= \frac{\sum y_i}{(\log \psi)^2} \frac{1}{\psi^2} - \frac{1}{\psi^2} \left( \frac{\sum y_i}{-\log \psi} - n \right) = \frac{1}{\psi^2} \left( \frac{\sum y_i}{(\log \psi)^2} + \frac{\sum y_i}{\log \psi} + n \right), \\ i^\Psi(\psi) &= \frac{n}{(-\log \psi)\psi^2}. \end{aligned}$$

Note that  $\hat{\psi} = e^{-\hat{\theta}}$ . △

## 1.5 Examples of Likelihood Functions

In this section some examples are given of likelihood functions for parametric statistical models which are different, in some way or another, from the more elementary cases of independent, identically distributed observations which, we presume, the reader is already familiar with.

### Example 1.2 Censored observations

Suppose  $T_1, \dots, T_n$  are independent and identically distributed continuous r.v.'s with density  $p_T(t; \theta)$ ,  $t \in \mathcal{T} \subseteq \mathbb{R}^+$ ,  $\theta \in \Theta$ , and distribution function  $F_T(t; \theta)$ . Let  $c_1, \dots, c_n$  be given positive constants. Suppose that the  $i$ -th observation ( $i = 1, \dots, n$ ) is a realization of the r.v.  $(Y_i, \delta_i)$ , with  $Y_i = \min(T_i, c_i)$  and  $\delta_i = I_{[0, c_i]}(T_i)$ , where  $I_A(\cdot)$  denotes the indicator function of the set  $A$ ,

defined by  $I_A(x) = 1$  if  $x \in A$ ,  $I_A(x) = 0$  if  $x \notin A$ . This observation scheme is called **type I censoring** (see e.g. Lawless, 1982, section 1.4) and is important in survival analysis. In this context,  $T_i$  denotes the survival time of the  $i$ -th unit and  $c_i$  is a preset censoring time for the same unit. Once the time  $c_i$  has elapsed, the unit is no longer observed and, if it is still alive its survival time is censored, that is, the only observation made is that  $T_i$  is greater than  $c_i$ . If  $\delta_i = 1$ , then  $Y_i$  is an uncensored lifetime; if  $\delta_i = 0$ , then the information on the  $i$ -th unit is that  $T_i > c_i$ . The r.v.  $(Y_i, \delta_i)$  has a mixed distribution, that is, it has one continuous and one discrete component, with joint density

$$p_{Y_i, \delta_i}(y_i, \delta_i; \theta) = \{p_T(y_i; \theta)\}^{\delta_i} \{1 - F_T(y_i; \theta)\}^{1-\delta_i}.$$

The likelihood function is

$$L(\theta) = \prod_{i=1}^n \{p_T(y_i; \theta)\}^{\delta_i} \{1 - F_T(y_i; \theta)\}^{1-\delta_i}. \quad (1.22)$$

If, for example,  $T_i$  has an exponential distribution with mean  $\theta$ , (1.22) becomes

$$L(\theta) = \theta^{-r} \exp(-\theta^{-1} \sum_{i=1}^n y_i),$$

where  $r = \sum_{i=1}^n \delta_i$  is the observed number of uncensored lifetimes.

If the censoring times  $c_i$  are independent realizations of a continuous r.v. with density  $p_C(c)$  and distribution function  $F_C(c)$  and if  $T_i$  and  $C_i$  are independent, then  $(Y_i, \delta_i)$  has density

$$p_{Y_i, \delta_i}(y_i, \delta_i; \theta) = \{p_T(y_i; \theta)(1 - F_C(y_i))\}^{\delta_i} \{p_C(y_i)(1 - F_T(y_i; \theta))\}^{1-\delta_i}. \quad (1.23)$$

Whenever the distribution of  $C_i$  does not depend on  $\theta$ , and this is a special case of **uninformative censoring**, the likelihood function based on  $(y_i, \delta_i)$ ,  $i = 1, \dots, n$ , is still given by (1.22). The same likelihood (1.22) is also obtained under weaker assumptions on the dependence of the survival and censoring mechanisms (see, for example, Lawless, 1982, section 1.4.1d).  $\triangle$

### Example 1.3 *Two-state Markov chain*

Let  $y = (y_1, \dots, y_n)$  be a realization of  $(Y_1, \dots, Y_n)$  whose p.d.f. is factorized as

$$p_Y(y) = p_{Y_1}(y_1) \prod_{i=2}^n p_{Y_i | Y_{i-1} = y_{i-1}}(y_i; y_{i-1}).$$

This holds if

$$p_{Y_i | Y_{i-1}=y_{i-1}, \dots, Y_1=y_1}(y_i; y_{i-1}, \dots, y_1) = p_{Y_i | Y_{i-1}=y_{i-1}}(y_i; y_{i-1}),$$

that is, if the dependence between the observations (which for ease of interpretation can be envisaged as a time series) is **Markovian**.

Consider the simplest case, where each  $Y_i$  can only take the values 0 and 1. In other words, we observe dependent binary variables which follow a first-order Markov chain. Assuming that the initial state  $y_1$  is fixed, the distribution of the observations for  $i = 2, \dots, n$  is entirely specified by the one-step transition probabilities, i.e. by the conditional probabilities

$$P(Y_i = 1 | Y_{i-1} = 0) = \theta_{01}$$

and

$$P(Y_i = 1 | Y_{i-1} = 1) = \theta_{11}.$$

Thus

$$p_Y(y) = \prod_{i=2}^n p_{Y_i | Y_{i-1}=y_{i-1}}(y_i; y_{i-1}) = \theta_{00}^{n_{00}} \theta_{01}^{n_{01}} \theta_{10}^{n_{10}} \theta_{11}^{n_{11}},$$

where  $\theta_{00} = 1 - \theta_{01}$ ,  $\theta_{10} = 1 - \theta_{11}$  and  $n_{jk}$ , ( $j, k = 0, 1$ ) denote the overall number of one-step transitions from state  $j$  to state  $k$ . Therefore, the likelihood function for  $\theta = (\theta_{01}, \theta_{11})$  is

$$L(\theta; y) = \prod_{j,k} \theta_{jk}^{n_{jk}}, \quad (1.24)$$

and this expression generalizes straightforwardly to Markov chains with more than two states. △

#### **Example 1.4** *Non-homogeneous Poisson process*

Poisson processes are special counting processes, i.e., continuous time stochastic processes with discrete state space. A counting process is a collection of random variables  $\{N_t, t \geq 0\}$ , where  $N_t$  can take the values  $0, 1, \dots$  and will express the number of arrivals or events that take place within the interval  $[0, t)$ . Let  $N_0 = 0$ . Indicate with  $N(t, t+h)$  the number of arrivals within the interval  $[t, t+h)$ , with  $t \geq 0$  and  $h > 0$ ; assume that  $N_t$  and  $N(t, t+h)$  are independent, that is, that the counting process has independent increments. It is said that  $\{N_t, t \geq 0\}$  is a **non-homogeneous Poisson process** if, as  $h \rightarrow 0$ ,

$$P\{N(t, t+h) = 0\} = 1 - \lambda(t)h + o(h)$$

and

$$P \{N(t, t+h) = 1\} = \lambda(t)h + o(h),$$

where  $\lambda(t)$  is a positive function called the **rate function** of the process, and  $o(h) \rightarrow 0$  as  $h \rightarrow 0$ . It can be shown that  $N(s, s+t)$  follows a Poisson distribution with mean  $\mu(s, t) = \int_s^{s+t} \lambda(u)du$ . The process is called **homogeneous** if  $\lambda(t) = \lambda$ ; in this case the times between two successive arrivals are independent and exponentially distributed with mean  $1/\lambda$ .

Suppose that the data correspond to the observation of a non-homogeneous Poisson process in the interval  $[0, t_0)$  and that the events take place at times  $t_1, \dots, t_n$ . One simple procedure for defining the likelihood function associated with these data, consists of subdividing the observation interval into  $m$  subintervals each  $h = t_0/m$  long, then, calculating the contribution (multiplicative, because of the independence of the increments) of each of the subintervals to the likelihood and, lastly, taking the limit of the factorization obtained as  $h \rightarrow 0$ . Denote these subintervals as  $[u_j, u_j + h)$ ,  $j = 1, \dots, m$ . The observation related to  $[u_j, u_j + h)$  contributes to the likelihood with a factor  $\lambda(u_j)h + o(h) = \lambda(t_i)h + o(h)$  if  $u_j \leq t_i < u_j + h$  for some  $i$ , that is, if there is one arrival in the interval, and with a factor of  $1 - \lambda(u_j)h + o(h)$  if in this interval there has been no arrival. Thus the likelihood has to be obtained from

$$\prod_{i=1}^n \{\lambda(t_i)h + o(h)\} \prod_j^* \{1 - \lambda(u_j)h + o(h)\}, \quad (1.25)$$

where the second product is over all values of  $j$  such that the interval  $[u_j, u_j + h)$  does not contain any arrival time  $t_1, \dots, t_n$ . Since

$$\begin{aligned} \prod_j^* \{1 - \lambda(u_j)h + o(h)\} &= \exp \left\{ \sum_j^* \log(1 - \lambda(u_j)h + o(h)) \right\} \\ &= \exp \left\{ - \sum_j^* (\lambda(u_j)h + o(h)) \right\}, \end{aligned}$$

the limit of this quantity as  $h \rightarrow 0$  is

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

Omitting the factor  $h^n$  in (1.25) we obtain

$$L(\lambda(t)) = \exp \left\{ - \int_0^{t_0} \lambda(u)du \right\} \prod_{i=1}^n \lambda(t_i). \quad (1.26)$$

The argument of  $L(\cdot)$  in (1.26) is a function, which is usually specified in a parametric class. If the process is homogeneous, i.e. with  $\lambda(t) = \lambda$ , (1.26) becomes

$$L(\lambda) = \lambda^n e^{-\lambda t_0}. \quad (1.27)$$

This likelihood is equivalent to that obtained regarding  $n$  as a realization of a Poisson distribution with mean  $\lambda t_0$ .

Extensions of the above considerations to spatial Poisson processes are particularly important in applications. Let  $A$  denote a bounded region of the plane and  $N(A)$  the number of arrivals in  $A$ . Then the function

$$\lambda(\mathbf{y}) = \lim_{|d\mathbf{y}| \rightarrow 0} \left\{ \frac{E(N(d\mathbf{y}))}{|d\mathbf{y}|} \right\},$$

where  $d\mathbf{y}$  denotes a neighbourhood of the point  $\mathbf{y} \in \mathbb{R}^2$  with area  $|d\mathbf{y}|$ , is called the **intensity function**. The collection  $\{N(A), A \subset \mathbb{R}^2, \text{bounded}\}$  is a non-homogeneous Poisson process in the plane if arrivals in disjoint regions are independent and if

$$P\{N(d\mathbf{y}) = 0\} = 1 - \lambda(\mathbf{y})|d\mathbf{y}| + o(|d\mathbf{y}|)$$

and

$$P\{N(d\mathbf{y}) = 1\} = \lambda(\mathbf{y})|d\mathbf{y}| + o(|d\mathbf{y}|).$$

Let

$$\Lambda(A) = \int_A \lambda(\mathbf{y}) d\mathbf{y}$$

denote the **intensity measure** of the process. By a repetition of the procedure leading to (1.26), the likelihood associated with the observation of points  $\mathbf{y}_1, \dots, \mathbf{y}_n$  in a fixed region  $A_0$  is seen to be

$$L(\lambda(\cdot); \mathbf{y}_1, \dots, \mathbf{y}_n) = \exp\{-\Lambda(A_0)\} \prod_{i=1}^n \lambda(\mathbf{y}_i).$$

△

### Example 1.5 *Brownian motion*

Brownian motion offers the most important example of a continuous time stochastic process with continuous state space. Brownian motion with drift coefficient  $\mu$  and unit variance is a collection of random variables  $\{Y(t), 0 \leq t < +\infty\}$  such that

- (i)  $Y(0) = 0$ ;
- (ii)  $P_\mu\{Y(t) - Y(s) \leq z\} = \Phi((z - (t - s)\mu)/(t - s)^{1/2})$ ,  $0 \leq s < t < +\infty$ ;
- (iii) for every finite sequence of points  $0 \leq s_1 < t_1 \leq s_2 < t_2 \leq \dots \leq s_n < t_n < +\infty$ ,  $n = 2, 3, \dots$ , the r.v.'s  $Y(t_i) - Y(s_i)$ ,  $i = 1, \dots, n$ , are independent;
- (iv)  $Y(t)$ ,  $0 \leq t < +\infty$ , is a continuous function of  $t$ .

In (ii)  $\Phi(\cdot)$  denotes the standard normal distribution function. Brownian motion is thus a process with normally distributed, independent increments.

Assume that the data correspond to the observation of Brownian motion in the interval  $[0, t_0]$ , that is, the partial trajectory  $\{y(t), 0 \leq t \leq t_0\}$  is observed. As in the preceding Example 1.4, a simple procedure for constructing a likelihood function is that of subdividing the observation interval into  $m$  subintervals  $h = t_0/m$  long, then calculating the multiplicative contribution (for independence of increments) of each subinterval to the likelihood and, lastly, by considering the limit as  $h \rightarrow 0$  of the factorization obtained. Let  $[u_j, u_j + h)$ ,  $j = 1, \dots, m$ , denote the subintervals. The observation related to  $[u_j, u_j + h)$  contributes to the likelihood with a factor

$$\exp \left\{ -\frac{1}{2h} \{y(u_j + h) - y(u_j) - \mu h\}^2 \right\},$$

proportional to

$$\exp \left\{ -\frac{1}{2} \mu^2 h + \mu \{y(u_j + h) - y(u_j)\} \right\}.$$

Hence, the overall likelihood is

$$L(\mu) = \exp \left\{ -\frac{1}{2} t_0 \mu^2 + \mu y(t_0) \right\}, \quad (1.28)$$

which is equivalent to the likelihood obtained regarding  $y(t_0)$  as a realization of a normal distribution with mean  $\mu t_0$  and variance  $t_0$ . Note that obtaining (1.28) does not require special limiting considerations. However, in more complicated cases, stochastic calculus is essential. From a general point of view, defining a likelihood requires it to be possible to express the probability measures associated with the process as a dominated family (cf. section 1.4.1). One technical difficulty that may arise is that the measures associated with processes with continuous sample paths could not be mutually absolutely continuous.  $\triangle$

## 1.6 Bibliographic Note

A large part of the material in this chapter is usually dealt with in introductory texts to the theory of statistical inference (see for example, Chapters 1 and 2 in Cox and Hinkley, 1974). General discussions of models, principles and methods can be found, among others, in Dawid (1983a), Fraser (1983), Cox (1958, 1986).

Other approaches to data analysis either supplement or oppose that of the theory of statistical inference which is based on the fundamental assumption that the data  $y^{obs}$  are a realization of  $Y \sim p^0(y)$ . Some views emphasize exploratory elements, description or classification. The expression *exploratory data analysis* is confined to all the preliminary techniques of data analysis suggested by Tukey (1977); see also Tukey (1980) and Chatfield (1985). The *analyse des données*, which was initially developed in France in the early 1960s, rejects the interpretation of data through probability models (see, for example, Benzécri, 1973, section TIA no:2) in favour of algebraic-geometric techniques of reduction and classification. Other possible non-probabilistic models that can be used to describe variability or uncertainty should also be mentioned. The most recent, and outstanding, example is given by the theory of chaotic deterministic systems. These have been discussed by Bartlett (1990), Berliner (1992) and Chatterjee and Yilmaz (1992). One final example is that of the theory of fuzzy sets, which aims to explain uncertainty as possibility: see, for example, Kaufmann and Gupta (1991).

For a lively resumé of the debate on the foundations of statistical inference, see Barnett (1982). The idea that because even the foundations of mathematics embody intricacies and elements of incompleteness, far more difficult problems are encountered when seeking to define general rules for statistical inference, has been repeatedly stressed by Barnard (see e.g. Barnard, 1974a). The distinction according to four visions of section 1.2 follows Cox (1993, 1995).

A classic text on the personalistic Bayesian paradigm is de Finetti (1974, 1975). Accounts of inference from the personalistic and non-personalistic Bayesian points of view may be found in the volumes by Berger (1985) and Bernardo and Smith (1994). Gelman, Carlin, Stern and Rubin (1995) give an introduction to Bayesian modelling.

For a fuller account of the Fisherian paradigm, see the appreciation of Fisher's contribution in Fienberg and Hinkley (1980), and, naturally, Fisher's collected works, Fisher (1950, 1971), and his last book, Fisher (1956). Rao (1992) summarizes Fisher's contribution to the development of statistics. For-

malization and extensions form the nucleus of the first part of Barndorff-Nielsen (1978). Cox and Hinkley (1974) offer a balanced introduction to the theory of statistical inference also with reference to other paradigms. The recent book by Lindsey (1996) stresses the central role of likelihood in frequentist as well as Bayesian approaches to statistical inference. The main references on the theory of inference, from the decisional point of view, are the classic books by Lehmann (1983, 1986) and Ferguson (1967). Efron (1986a) offers other points for reflection in favour of the adoption of frequentist paradigms within scientific research. McPherson (1989) and Hand (1994) discuss the interaction between statistics and its users in various scientific fields.

Useful hints on model specification are given in Lehmann (1990) and Cox (1990); see also Cox and Hinkley (1974, sections 1.2 and 1.3). Chatfield (1995) and Draper (1995) discuss model uncertainty. Introductions to statistical inference in the context of nonparametric models, with emphasis on different aspects, can be found in Randles and Wolfe (1979), Tarter and Lock (1994), Efron and Tibshirani (1993). Pfanzagl (1990) and Bickel, Klaassen, Ritov, Wellner (1993) are recent monographs on semiparametric models. As an example of parametric statistical models which explain the phenomenon being studied, see Barndorff-Nielsen, Blæsild, Jensen and Sørensen (1985) where a parametric model which aims to describe the variability of the mass of grains of sand is derived. See Tawn (1990) for one use of stable distributions as an element for modelling unobservable quantities. For a more in-depth examination of methods for selecting a parametric model, goodness-of-fit tests and analysis of residuals, see Linhart and Zucchini (1986), D'Agostino and Stephens (1986), Cook and Weisberg (1982).

For a detailed introduction to the notions of measure theory mentioned in section 1.4.1, a useful reference is for instance Billingsley (1986). The notation in sections 1.4.2 and 1.4.3 is introduced in Barndorff-Nielsen (1988). The concept of likelihood appears in the first scientific paper published by Fisher, in 1912, when he had just taken his honours degree in mathematics at Cambridge University. Fisher's early works emphasized likelihood as a tool for tracing estimators which, in many ways, are preferable to those based on the method of moments. He subsequently developed the theory of likelihood in papers published in 1921, 1922a, 1925 and 1934, and in the book published in 1956. From the historical point of view, the idea of estimates based on the maximization of  $L(\theta)$  is usually attributed to Fisher (1912, 1922a, 1925, 1935), however, an embryonic version of the idea had already been put forward by Lambert, Daniel Bernoulli and Lagrange in the eighteenth century, see Edwards (1974)

for a lively account. Recent extensions of the theory to semiparametric and nonparametric models are dealt with e.g. in Gill (1989), Wong and Severini (1991), Gill and van der Vaart (1993), Lindsay (1995), Murphy (1995a), van der Vaart (1996). There are many examples of the application of the method of maximum likelihood to problems of estimation in complex models, to cite just a few: Vardi and Lee (1993), Ying (1993), Coles and Tawn (1994).

The examples offered in section 1.5 are common. See Cox and Snell (1989, section 2.11) for a more in-depth treatment of two-state Markov chains and Bhat (1985) and the references given there for more complex Markov processes. The procedure of constructing the likelihood function used in Examples 1.4 and 1.5 is adopted in Cox and Hinkley (1974, p. 15), Barndorff-Nielsen (1991a, section 10.2.6), Barndorff-Nielsen and Cox (1984a). Further examples of the technique for calculating the likelihood function for a stochastic process with a continuous parameter can be found in Barndorff-Nielsen (1991a) and in Barndorff-Nielsen and Cox (1994, section 2.2.3). A rigorous formalization of the procedure adopted could be based on the notion of **product integral** (Gill and Johansen, 1990); see also Andersen, Borgan, Gill and Keiding (1993, Chapter 2). For a definition of the likelihood function related to a general class of processes, the **diffusion processes**, see Sørensen (1983) and Barndorff-Nielsen and Sørensen (1994) where various examples are presented.

## 1.7 Problems

**1.1** Justify the statement that, for a discrete r.v. with finite support, any statistical model is a parametric one.

[Section 1.3.1]

**1.2** Consider the parametric model  $(\mathcal{Y}, P_\theta, \Theta)$ , with  $\mathcal{Y} = [0, 1]$ ,  $\Theta = \{\theta_1, \theta_2\}$ , where  $P_{\theta_1}$  is the measure that corresponds to the uniform distribution on  $[0, 1]$ , while  $P_{\theta_2}$  is the measure that corresponds to the discrete distribution taking values  $1/2$  and  $1$  with equal probability. Show that the probability measures  $P_{\theta_1}$  and  $P_{\theta_2}$  are not mutually absolutely continuous. Show that the parametric family defined above is dominated by the measure  $P_{\theta_1} + P_{\theta_2}$ .

[Section 1.4.1]

**1.3** A subject undergoes  $n$  tests of ability. Let  $Y_i$ ,  $i = 1, \dots, n$ , be the result of the  $i$ -th test; put  $Y_i = 1$  if the answer is correct and  $Y_i = 0$  if the answer is wrong. Assume that the probability of a correct answer at the  $i$ -th test is  $\pi_i(\theta) = \exp(\theta - x_i) / \{1 + \exp(\theta - x_i)\}$ ,  $\theta \in \mathbb{R}$ , with  $x_1, \dots, x_n$  given real

constants. Furthermore, assume that  $Y_1, \dots, Y_n$  are independent. Specify the statistical model for  $Y_1, \dots, Y_n$ . Generally, the parameter  $\theta$  and the constant  $x_i$  are referred to, respectively, as *ability of the subject* and *difficulty of the  $i$ -th test*; justify this interpretation. Write the likelihood equation for  $\theta$  and say when it has a finite solution.

[Section 1.4]

**1.4** Suppose  $y_1, \dots, y_n$  are i.i.d. observations from an exponential distribution with mean  $1/\lambda$ ,  $\lambda > 0$ . Only the integer part is observed for the last  $n - m$  ( $m < n$ ) observations. Specify the statistical model. Write the likelihood equation for  $\lambda$  and show that this has only one solution.

[Section 1.4]

**1.5** Prove and interpret the relation  $\hat{J}_{ab} = \sum_{r,s} \hat{J}_{rs} \hat{\theta}_a^r \hat{\theta}_b^s$ .

[Section 1.4.3]

**1.6** Let  $y_1, \dots, y_n$  be  $n$  i.i.d. observations from a normal distribution with mean  $\mu$  and variance  $\sigma^2$ . Obtain  $l_*$ ,  $i$  and  $\hat{j}$  in the parameterizations  $\theta = (\mu, \sigma^2)$  and  $\psi = (\mu, \sigma)$ .

[Section 1.4.3]

**1.7** Let  $y_1, \dots, y_n$  be  $n$  i.i.d. observations from a Poisson distribution with mean  $\mu$ . Obtain  $l_*$ ,  $i$  and  $\hat{j}$  in the parameterizations  $\theta = \mu$  and  $\psi = \mu^\delta$ , with  $\delta \neq 0$  fixed.

[Section 1.4.3]

**1.8** Let  $y_1, \dots, y_n$  be  $n$  independent observations from a binomial distribution with index  $m$  and parameter  $\theta$ ,  $Bi(m, \theta)$ . Obtain  $l_*$ ,  $i$  and  $\hat{j}$  in the parameterization  $\psi = \Phi^{-1}(\theta)$ , where  $\Phi(\cdot)$  indicates the standard normal d.f., starting from the same quantities calculated in the parameterization  $\theta$ , and using the transformation laws (1.13), (1.19) and (1.20).

[Section 1.4.3]

**1.9** Let  $y_1, \dots, y_n$  be independent observations from a two-parameter gamma distribution, with p.d.f.  $p(y; \nu, \lambda) = \lambda(\lambda y)^{\nu-1} e^{-\lambda y} / \Gamma(\nu)$ , with  $y > 0$ ,  $\nu > 0$ ,  $\lambda > 0$ . Obtain  $l_*$ ,  $i$  and  $\hat{j}$  in the parameterizations  $\theta = (\nu, \lambda)$  and  $\psi = (\nu, \nu/\lambda)$ .

[Section 1.4.3]

**1.10** Let  $y_1, \dots, y_n$  be independent observations from a two-parameter Weibull distribution with p.d.f.

$$p(y; \nu, \lambda) = \lambda \nu (\lambda y)^{\nu-1} e^{-(\lambda y)^\nu}, \quad (1.29)$$

with  $y > 0$ ,  $\nu > 0$ ,  $\lambda > 0$ . Obtain  $l_*$ ,  $i$  and  $\hat{j}$  in parameterizations  $\theta = (\nu, \lambda)$  and  $\psi = (\nu, \lambda^\nu)$ .

[Section 1.4.3]

**1.11** Write the likelihood function (1.24) according to the alternative parameterization  $\psi = (\beta_0, \beta_1)$ , with  $\theta_{01} = e^{\beta_0}/(1 + e^{\beta_0})$ ,  $\theta_{11} = e^{\beta_0 + \beta_1}/(1 + e^{\beta_0 + \beta_1})$ . Interpret the parameter  $\beta_1$ .

[Section 1.5]

**1.12** Let  $y_0, y_1, \dots, y_n$  be realizations of  $n + 1$  random variables  $Y_0, Y_1, \dots, Y_n$  such that:

(i)  $Y_0 \sim N(\mu, \sigma^2)$ ;

(ii) The distribution of  $Y_i$  given  $Y_0 = y_0, \dots, Y_{i-1} = y_{i-1}$  is the same as that of  $Y_i$  given  $Y_{i-1} = y_{i-1}$  and is normal with mean  $\mu$  and variance  $\sigma^2(1 + y_{i-1}^2)$ ,  $i = 2, \dots, n$ .

Write the likelihood function for  $(\mu, \sigma^2)$  and the corresponding likelihood equations. Obtain the maximum likelihood estimate for  $(\mu, \sigma^2)$  and suggest an interpretation of the formula obtained for  $\hat{\mu}$ .

[Section 1.5]

**1.13** Specialize the likelihood function (1.26) when  $\lambda(t) = \alpha + \beta t$  and when  $\lambda(t) = \alpha t^\beta$ .

[Section 1.5]

**1.14** Suppose  $\{N_y, y \geq 0\}$  is a non-homogeneous Poisson process with rate function

$$\lambda(y; \xi, \sigma, \mu) = \sigma^{-1} \{1 + \xi(y - \mu)/\sigma\}^{-1/\xi - 1},$$

if  $1 + \xi(y - \mu)/\sigma > 0$  and zero elsewhere, with  $\xi, \mu \in \mathbb{R}$ ,  $\sigma > 0$ . For  $\xi = 0$  the rate function  $\lambda(y; 0, \sigma, \mu)$  is defined as  $\lim_{\xi \rightarrow 0} \lambda(y; \xi, \sigma, \mu) = \exp\{-(y - \mu)/\sigma\}$ . Arrivals are observed at  $y_1, \dots, y_m$ . Using (1.26), write the likelihood function for  $(\xi, \mu, \sigma)$ . (*Hint*: the observations must fall within  $I = \{y > 0 : 1 + \xi(y - \mu)/\sigma > 0\}$ .) This model is used for the analysis of extremes.

[Section 1.5; Smith, 1989, section 4]