CHAPTER 1

Introduction

Statistics is concerned with *information*. This is what the subject of this book is all about. A problem in statistics begins with an unknown parameter θ, i.e. a variable whose exact value is not known. In order to generate information about θ we perform a statistical experiment $\mathcal{E}$, which generates data x. The problem is then to extract the maximum information about the unknown parameter θ from the experiment $\mathcal{E}$ and the data x. This is called *statistical inference*. But what is this information? And how is the statistical inference performed? Different competing schools of thought have developed over the course of time. The most popular ones are Bayesian statistics, Neyman-Pearson theory and Fisher theory and its different variants. Although these schools seem to be irreconcilable, they agree to some extent on the formal elements needed for statistical inference.

To start with, we have to ask how an *experiment* $\mathcal{E}$ is described. There exists a common mathematical framework widely used in statistics for the notion of a statistical experiment: the set X, the *sample space*, is the set of all possible outcomes (samples) of the experiment. The set Θ, the *parameter space*, is the set of all possible values for the unknown parameter θ. For the sake of simplicity we assume X and Θ to be *finite*. Finally, $p_\theta(x)$ is a family of probability measures on the sample space X, indexed over all possible values $\theta \in \Theta$.

In the Bayesian school a *prior* probability distribution $p_0(\theta)$ over the parameter space Θ is further assumed to be given. The measure $p_\theta(x)$ is then interpreted as a conditional probability and written as $p(x|\theta)$ to emphasize this point of view. The whole setup can then be summarized in the joint probability distribution

$$p(x, \theta) = p(x|\theta)p_0(\theta).$$

Then the experiment is performed and data x obtained. The information about the unknown parameter θ can be summarized by the conditional probability of θ given x,

$$p(\theta|x) = c \cdot p(x|\theta)p_0(\theta),$$

where c is a normalization constant given by

$$c^{-1} = p(x) = \sum_{\theta \in \Theta} p(x|\theta)p_0(\theta).$$

This is Bayes' theorem. This is a clear and neat approach. The only problem is that there are situations where it is difficult to assume that a prior distribution $p_0(\theta)$ is known. In particular there might be complete ignorance about θ. Bayesians then usually assume a uniform distribution over Θ. But this is rather a

default assumption, which replaces ignorance. Also, if Θ is infinite such a uniform distribution does not exist. Finally, if the goal is to infer about some other parameter $\lambda = g(\theta)$, where the function g is known, then a uniform distribution over θ does not, in general, transform into a uniform distribution over λ. So there are different ways to represent ignorance by probability distributions, and it is by no means clear which is the "correct" one. These are a few of the objections against Bayesian inference. This is why other schools of statistical inference prefer to avoid the use of prior distributions.

As an alternative, the *likelihood principle* [**1**] states that all the statistical information is contained in the likelihood function

$$l(\theta) = p_\theta(x),$$

where on the right hand side x is the observed data which is fixed now. See [**2**] for a discussion of the likelihood paradigm. Here no prior distribution is used. In fact, it is not clear whether and how prior information can be added in this framework. However, the likelihood principle is related to Fisher's concept of fiducial distribution through the notions of sufficient and ancillary statistics. Many controversies arose out of these concepts and we do not want to go into the details of these discussions. See Hampel [**3**] for a recent perspective on the fiducial argument and fiducial probabilities.

However, we claim and show that the likelihood function alone cannot, in general, be considered to carry the full statistical information. Yet, we are still interested in inference methods which do not necessarily rely on a prior distribution on the unknown parameter. As in the Bayesian approach and Fisher's fiducial inference we want to reason towards "posterior" probability statements for the unknown parameter. But unlike Fisher's fiducial inference and unlike Neyman-Pearson theory, we want to integrate prior information into the inference, if it is available. More generally, we want to integrate any additional information that is available. Also unlike Bayesian inference, we want to integrate prior information which is not necessarily given in the form of a prior probability distribution. We claim that this is possible with an approach containing Bayesian inference as a special case and which reproduces Fisher's fiducial distributions, which get a clear meaning in the theory developed in this book.

We propose to describe statistical experiments through *functional models*, which define how data x is generated from the unknown parameter and some stochastic elements. The use of functional models is of course in itself nothing new. Dawid, Stone, and Bunke, among others, have used this type of models to explain fiducial inference [**4, 5**]. Another author, Fraser, based his approach of inference on structural models, a variant of functional models [**6**]. For example, least squares methods of regression analysis and Kalman filtering are based on functional models. However, our use of functional models is different: we base the inference from functional models and related observed data on the principle of *assumption-based reasoning*. This allows us to make "probabilistic" statements

about the unknown parameter that have a clear meaning. In addition, our approach is not limited to the special class of *invertible functional models*, which have been exclusively considered in the literature so far (except [**7, 8**]). The approach presented in this book is much more general.

The benefits of assumption-based reasoning based on functional models are multiple:

(1) Like Bayesian inference, the reasoning is towards posterior probabilistic information about the unknown parameter, but without the necessity of a prior probability distribution.
(2) The method allows the inclusion of prior or otherwise additional information about the unknown parameter, if available. While this information can be given by a prior probability distribution, other and more general forms of information are possible.
(3) Assumption-based inference on functional models subsumes both Bayesian and fiducial inference as special cases.
(4) Although unknown parameters are not considered as random variables, which they are definitely not, the approach gives a clear meaning to probabilistic statements about the unknown parameter in the form of probabilities of provability or derivability of hypotheses about the parameter.

Probabilistic assumption-based reasoning has been used by Pearl [**9**] to explain belief functions in the sense of Dempster-Shafer theory [**10**] in a probabilistic framework. This kind of reasoning has been developed so far especially in the context of propositional logic [**11, 7, 12**]. A first discussion of its application to statistical problems, especially to inference in linear systems with normal disturbances, is given in [**8**].

In a series of important papers Arthur Dempster generalized Bayesian inference and proposed new methods for reasoning towards posterior distributions based on samples [**13, 14, 15, 16**]. The goal was to show that an inference in the spirit of Bayesian inference is also possible without a prior distribution. In this way, the approach of Fisher based on fiducial probabilities and Bayes' approach would be reconciled as special cases of a unifying formalism. Fisher's approach corresponds to complete ignorance about the parameter before sampling, whereas Bayes' inference already assumes a complete prior information. In Dempster's approach a whole spectrum of intermediate prior information between these two extremes is available. Our method is closely related to this work, but with an alternative, new look at the underlying formalism. Dempster introduced lower and upper probabilities on hypotheses about the unknown parameter. In assumption-based reasoning, the lower probability becomes the probability of derivability of a hypothesis and the upper probability becomes the probability of non-derivability of the negation of the hypothesis, where derivability is investigated from the given functional model and the data.

Dempster's method of upper and lower probabilities was the motivation for G. Shafer to develop a mathematical theory of evidence [**10**]. This became the basic

source for the development of the so-called Dempster-Shafer theory of evidence. However, Shafer considered belief as an epistemological measure of uncertainty. Despite the formal similarity of the resulting statements, this interpretation is quite different from the interpretation of statements coming from probabilistic assumption-based reasoning. Nevertheless, Shafer also discussed statistical inference using evidence theory [**17**]. Another author, P. Smets, studied statistical inference based on belief functions with yet another, non probabilistic, semantics [**18, 19, 20, 21**].

In the *first part* of this book we examine *discrete* functional models, where both the sample space and the parameter space are finite sets. This simple case of discrete models is used to present the basic ideas of assumption-based inference in a form which is unhampered with technical difficulties.

In the case of discrete models, the theory can be given an elegant form based on the theory of hints [**7, 22**]. This form emphasizes that statistical inference is concerned with the combination of several pieces of information. The combined information is then focussed to the relevant aspects or questions. This gives an *algebraic* flavor to the inference process (see Chapter 4). This perspective also shows that assumption-based inference very naturally leads to familiar concepts in the Dempster-Shafer theory of evidence [**10**]. It also gives a clear semantics to the classical notion of likelihood function and clarifies its true place and role in the field of statistical inference. In particular, it will be shown that the likelihood function does not, in general, represent the entire information in a statistical experiment. However, in some special situations, it does represent the entire information. A necessary and sufficient condition for this to happen will be presented (Section 3.4).

The first part also discusses how additional information coming from outside the experiment or from other experiments can be integrated into the inference process. This includes the Bayesian case, where prior information is assumed to be given in the form of a probability distribution on the unknown parameter. It will be shown that, in general, prior ignorance about the parameter *cannot* be represented by a uniform distribution over the parameter space (Section 3.3). Again, there are particular situations where prior ignorance can safely be represented by a uniform distribution. It will be shown that this is the case when the functional model is invertible (Section 3.4).

Finally the question of how to decide which hypotheses are credible in view of the available information is addressed in Chapter 5. A simple rule called the α-rule is proposed and its properties are examined. In particular, the risks of errors in accepting or refuting hypotheses are elucidated. An important feature of assumption-based inference is to consider that it is not reasonable to require that a hypothesis must always be accepted or rejected. Note that this point of view is also shared by the Dempster-Shafer theory. In some cases the decision must be left open. For example, in the extreme case of total ignorance, it does not make sense to accept or reject a hypothesis because there is simply not enough information to decide. This leads to the question of how does the procedure specified

by the α-rule compare to the classical Neyman-Pearson tests theory. This is discussed in Section 5.2. In particular, the conceptual difference between the *seller's* position underlying Neyman-Pearson's tests theory, which controls average error probabilities over the whole sample space (the operational characteristics of the test), and the *user's* position, which is to control the errors *given* the observed sample, is discussed. Assumption-based reasoning, like Bayesian inference, takes the second position. Nevertheless, it is shown that in some cases the α-rule coincides with most powerful Neyman-Pearson tests. The general consistency of the α-rule is also proved. The difference between the seller's and buyer's position has also been discussed by Hacking in terms of "before-trial betting" and "after-trial betting" [**23**].

One must be careful in generalizing the results from the discrete to the continuous case. Therefore, in the *second part* of the book we treat several variants of continuous models. For these models the theory is far less complete than for discrete models. For this reason, part II concentrates on assumption-based inference in invertible models only. Nevertheless, this is already sufficient to review and clarify several results and difficulties of Fisher's fiducial theory and the likelihood principle. In the same spirit, this also permits to examine the role of improper priors in Bayesian inference. In fact, they get a clear meaning as elements of the Theory of Hints and the Dempster-Shafer Theory (Section 9.2).

In Chapter 7 the basic cases of simple and partitionable models are presented. The concept of partitionable models is due to Dawid, Stone and Bunke [**4, 5**]. In both cases *fiducial probability measures* on the parameter space can be deduced. These are particular support functions in the sense of the theory of hints, or particular belief functions in the sense of the Dempster-Shafer theory of evidence. Important special cases of partitionable models are linear and *structural* models. The latter concept is due to Fraser [**6, 24**]. These cases are examined in Chapter 8. Structural models lead to reduced models and sometimes *sufficient reduced* models. It is conjectured that the assumption-based concept of sufficiency is identical to Fisher's classical concept (Section 9.1). This is verified in some examples, but not formally proved in general.

As in the discrete case, adding external information to the inference is easily possible (Section 9.2). In particular, different forms of prior knowledge, like for example total ignorance, restriction of the parameter to a subset of its space, uniform or other proper or improper prior densities can be integrated into the assumption-based inference. This clarifies greatly the role of proper and improper Bayesian priors, as well as other forms of additional information.

As in discrete models, it may be necessary to select hypotheses. The α-rule is again proposed (Chapter 10). In particular this leads to $(1-\alpha)$-*fiducial intervals*, which are shown to be different from classical *confidence intervals*. Like classical statistical tests, the latter are concepts developed from a *seller's* position. They guarantee overall error probabilities. Fiducial intervals on the other hand guarantee error probabilities *given* the observed sample, which reflects the *user's* position. But again in some cases the two concepts coincide and they

are closely related to the notion of sufficiency. It must be stressed, however, that the relations between assumption-based decision making and classical statistical decision making procedures (tests, confidence intervals, estimators) are far from being satisfactorily clarified. Much more research is needed in this direction.

The case of *linear models*, especially with Gaussian disturbances, is particularly important from a practical point of view. Therefore, the *third part* of the book is devoted to this subject. As an introduction, a simple economic model borrowed from Pearl [**25**] is discussed in Chapter 12. Then linear systems in general are treated (Chapter 13). But as is to be expected, the most interesting results are obtained for Gaussian disturbances (Chapter 14). As in the discrete case, it is possible to represent the information carried by linear systems with Gaussian disturbances by (Gaussian) hints (Chapter 15). This stresses once again the information aspect of statistical models and observations. Furthermore, as in the discrete case, Gaussian hints induce an algebraic structure which captures the operations of combining and focussing of information (Chapter 16). This structure is an instance of a valuation algebra [**26**], [**27**], which is a particular case of a more general algebraic theory providing a framework for computing with pieces of information [**28**]. This is illustrated by the famous *Kalman filter* (Section 16.5). A similar perspective on the Kalman filter has also been proposed by Dempster [**29**], [**30**]. The idea of normal belief functions presented in these papers is further developed in [**31**] and [**32**]. Normal belief functions are also called Gaussian belief functions or linear belief functions. An application of these functions in the field of finance is given in [**33**]. We also note that a method for analyzing dynamical systems with stochastic disturbances or with unknown but bounded disturbances is presented in [**34**]

As a final remark we would like to indicate that probabilistic assumption-based reasoning, as discussed in this book in relation to statistical inference, is also successfully applied in *artificial intelligence*, where it leads to a new way of combining logic with probability [**35, 12**]. Assumption-based reasoning provides a generic approach to reasoning under uncertainty, combining the two classical theories of inference, probability and logic [**28**]. In fact, this approach is not so new after all since traces of it can already be found in the famous *Ars Conjectandi* of Jakob Bernoulli [**36**].

Part 1

Discrete Models

CHAPTER 2

Inference with Functional Models

2.1. Introduction

An experiment $\mathcal{E}$ is usually described in statistics by the *sample space* X, the set of all possible outcomes of the experiment, the set Θ, the *parameter space*, the set of all possible values of an unknown parameter θ. Finally a parametric family of probability measures p_θ, $\theta \in \Theta$, on the sample space X is specified. When the experiment $\mathcal{E}$ is performed, an outcome $x \in X$ is obtained or observed. This is the *data*. From the data x and the description of the experiment $\mathcal{E}$ by the probability measures p_θ inference about the unknown parameter θ is attempted. Classically, this usually amounts to constructing an *estimator* or a *confidence interval* for the unknown parameter θ, or else to make a test on a pair of alternative hypotheses H_0 and H_1 about the parameter θ. Estimators, confidence intervals and test procedures are selected on the basis of the considerations of possible errors, trying to minimize somehow these errors. That is, statistical procedures are selected on the basis of their operational characteristics, i.e. on the probabilistic description of their average performances over the possible samples. The emphasis is on the *average* behavior of procedures. Little or nothing is known then on the behavior of the procedure in the individual case of an actual data x.

Bayesian statistics is different. Here, based on a prior probability distribution, the statistical model p_θ and the data x is used to determine a *posterior distribution* $p(\theta|x)$ of the unknown parameter θ, using Bayes' theorem. In this approach probabilistic statements about the unknown parameter can be made in the *individual case* of an actual data x. So emphasis is not on the average behavior of statistical procedures, but on the inference in each individual case. From the point of view of a user (a buyer) of statistical procedures, this is preferable because the risks taken in making decisions based on the inference are known in the individual case, and not only in the average.

The critical point however in Bayesian statistics is that a *prior distribution* must be used. This can be justified in the framework of a subjective probability theory. Nevertheless, in many practical cases it would be preferable to avoid the assumption of a prior distribution because there is no prior knowledge about the unknown parameter. But still it is desirable to make probabilistic statements about θ in the individual case of the actual data x. This has been attempted by Fisher [**37**] with the concept of *fiducial probability*. This subject was the object of many controversies. The meaning of fiducial probability itself was not altogether clear. Further, it is not always clear how to obtain these fiducial

probabilities, and several apparent paradoxes have been reported. The basic element used to obtain fiducial probabilities is the *likelihood function*, and it is supposed that this function contains all information that can be extracted from a statistical experiment. We claim that this is false in general, that it is the source of many difficulties and also the source of the problem of properly interpreting fiducial probabilities. Our claim is that we need *functional models* to properly deduce fiducial probabilities and that, in general, these models are not uniquely determined by likelihood functions.

The use of functional models for fiducial inference is not new [**6, 4, 5**]. But these authors used only a particular class of functional models, i.e. those which we call *invertible* (see Section 3.4 for the definition). So, although invertible models are important, they do not cover all cases. The use of non-invertible models, on the other hand, introduces the new problem that only bounds for "fiducial" probabilities can be obtained [**13, 14, 15, 16**]. We reconsider Dempster's approach using *assumption-based reasoning* to give a clear meaning to probabilistic statements about the unknown parameter in the case of individual, actual data x, i.e. fiducial probabilities, in all cases. In this way, a unified approach to statistical inference, including Fisher's fiducial method and Bayes' approach as special cases, is obtained. This method reproduces many well known results but also leads to new insights and methods. In particular, it clarifies the relations between Fisher's approach and Bayes' statistics, especially the role of priors, including improper priors (in particular in the continuous case). It also puts the likelihood function into its proper place and shows in which cases this function really contains the whole statistical information.

We introduce assumption-based reasoning with functional model in the case when both the sample space X as well as the parameter space Θ are *finite*. This allows the development of an elementary and elegant complete theory of statistical inference. However, care must be exercised in generalizing this elementary theory to more general cases of infinite sample spaces or infinite parameter spaces. These more general situations are discussed in the next part of the book.

2.2. Functional Models

Functional models describe how data x is generated from a parameter θ and some random element, designated by ω. We assume that the random element ω comes from some set Ω, which is also supposed to be *finite*. Let then f be a given function $f : \Theta \times \Omega \to X$ such that

$$x = f(\theta, \omega).$$

So, if a parameter value $\theta \in \Theta$ is given and a random element $\omega \in \Omega$ is selected, then data x is uniquely determined by the function f. In this sense, a functional model describes the process of data generation. We assume not only that f is given, but also that a probability measure P on Ω is known. This probability

measure is given by the probabilities $p(\omega)$ for all $\omega \in \Omega$ and of course,

$$p(\omega) > 0, \text{ for all } \omega \in \Omega, \quad \sum_{\omega \in \Omega} p(\omega) = 1.$$

We stress that these probabilities do not depend on θ. These elements, the function f and the probabilities p constitute a *functional model* for a statistical experiment $\mathcal{E}$.

Note that, if we assume a parameter θ, then, from the probabilities p, we can compute the probabilities for the data x,

$$p_\theta(x) = \sum_{\omega : x = f(\theta, \omega)} p(\omega).$$

Hence, a functional model induces a parametric family of probability measures on the sample space X, an object which is usually assumed a priori in modeling statistical experiments. We emphasize however that *different* functional models may induce the *same* parametric family $p_\theta(x)$ of probability measures. So, functional models contain *more* information than the family $p_\theta(x)$.

EXAMPLE 2.1 (A simple coin). Consider a fair coin with faces designated by 1 and 2. Suppose that there are only two possible cases: either face 1 carries heads and face 2 tails (case designated by parameter θ_0), or both faces carry heads (case designated by parameter θ_1). Thus we have $\Theta = \{\theta_0, \theta_1\}$. The observed outcome x of the experiment $\mathcal{E}$ consisting in throwing the coin once is either *heads* or *tails*. This means that $X = \{heads, tails\}$. The chance element ω finally is simply the face 1 or 2 turning up, $\Omega = \{1, 2\}$, with $p(1) = p(2) = 1/2$. The functional model is then completed by the function f defined as follows:

$$x = \begin{cases} heads & \text{if } \theta = \theta_0, \text{and } \omega = 1, \\ heads & \text{if } \theta = \theta_1, \\ tails & \text{if } \theta = \theta_0, \text{and } \omega = 2. \end{cases}$$

This functional model induces the following statistical specification:

$$p_{\theta_0}(heads) = p_{\theta_0}(tails) = \frac{1}{2},$$
$$p_{\theta_1}(heads) = 1, \quad p_{\theta_1}(tails) = 0.$$

△

EXAMPLE 2.2 (An urn model). Suppose an urn contains N balls numbered from 1 to N. The first θ balls are white, the rest is black, $0 \leq \theta \leq N$. The experiment $\mathcal{E}$ consists of drawing one ball from the urn and observing its color. The unknown parameter is θ. Then we have $X = \{black, white\}$ and $\Theta = \{0, 1, \ldots, N\}$. The random element is the number of the ball drawn, thus $\Omega = \{1, \ldots, N\}$, and we assume that $p(\omega) = 1/N$ for all ω. The functional model is then completed by

$$x = \begin{cases} white & \text{if } \omega \leq \theta, \\ black & \text{if } \omega > \theta. \end{cases}$$

The corresponding family of probabilities $p_\theta(x)$ is

$$p_\theta(white) = \frac{\theta}{N}, \quad p_\theta(black) = 1 - \frac{\theta}{N}.$$

Of course in this example, like the first one, not much information can be obtained from a single draw. The experiment must be repeated, either with or without replacement of the ball drawn. In the first case, with n draws, let $x = (x_1, \ldots, x_n) \in X^n$ and $\omega = \{\omega_1, \ldots, \omega_n) \in \Omega^n$ with $p(\omega) = 1/N^n$. Define then $x = f(\theta, \omega)$ component-wise for $i = 1, \ldots, n$ by

$$x_i = \begin{cases} white & \text{if } \omega_i \leq \theta, \\ black & \text{if } \omega_i > \theta. \end{cases}$$

This is then the functional model for n draws with replacement. △

Example 2.3 (A sensor model). Consider a sensor which should detect the presence of some hazardous material like smoke, gas, water, etc. Let θ_1 denote the presence of the hazardous material and θ_0 its absence. So $\Theta = \{\theta_0, \theta_1\}$. The chance element ω comes from the possible failure of the sensor to operate properly. So the sensor is either intact (i) or faulty (f), so that $\Omega = \{i, f\}$. It is assumed that probabilities of these two possible states are known to be $p(i) = p$ and $p(f) = 1-p$. Since the sensor may produce the alarm (a) or remain silent (s), the set of possible observations is $X = \{a, s\}$. Furthermore, it is supposed that an intact sensor correctly indicates the situation whereas a faulty sensor produces an alarm when there is no need and remains silent when hazardous material is present. The following functional model describes this situation

$$x = \begin{cases} a & \text{if } \theta = \theta_1, \omega = i \text{ or } \theta = \theta_0, \omega = f, \\ s & \text{if } \theta = \theta_0, \omega = i \text{ or } \theta = \theta_1, \omega = f. \end{cases}$$

More elaborate sensor models are possible [**7, 12**]. △

2.3. Assumption-Based Reasoning with Functional Models

Consider a functional model $x = f(\theta, \omega)$, with given probabilities $p(\omega)$ of the random elements, describing an experiment $\mathcal{E}$. Suppose that the outcome of the experiment is observed to be x. Given this data x and the experiment $\mathcal{E}$, what inference can be made about the value of the unknown parameter θ? The basic idea of assumption-based reasoning is to assume a random element ω and look what can be deduced for the unknown parameter under this assumption. Afterwards, the deductions are weighed by the probabilities of the unknown random elements. This will now be described in details.

First, note that by the observation x, some chance elements ω may become a posteriori impossible. In fact, if, for an $\omega \in \Omega$, there is no $\theta \in \Theta$ such that $x = f(\theta, \omega)$ holds, then this ω is clearly impossible: it cannot have generated the actual observation x. So, the observation x induces an event in Ω, which we call v_x,

$$v_x = \{\omega \in \Omega : \text{there is a } \theta \in \Theta \text{ such that } x = f(\theta, \omega)\}. \tag{2.1}$$

Since by the observation x, we know the event v_x has happened, we condition the probability measure P on Ω to this event v_x. This leads to the revised probabilities $p'(\omega) = p(\omega)/P(v_x)$ for all $\omega \in v_x$ (and $p'(\omega) = 0$ for $\omega \notin v_x$).

Within v_x it remains unknown which chance element ω caused the observation x, only their respective probabilities $p'(\omega)$ are known. Nevertheless, let us suppose for the time being that ω caused the observation x. Then, the possible parameter values θ can be restricted to the set

$$T_x(\omega) = \{\theta \in \Theta : x = f(\theta, \omega\}. \tag{2.2}$$

Note that in general it will contain several elements, although in some cases this is a one-element set. It is even possible that $T_x(\omega) = \Theta$, in which case the observation x, assuming ω, would carry no information about θ. So, in general, even if the chance element causing the observation would be known, this would not permit to identify the unknown parameter unambiguously.

But consider a hypothesis about the unknown parameter. Such a hypothesis is simply described by a subset $H \subseteq \Theta$. This hypothesis may be true or false, and again, the question can, in general, not be decided unambiguously on the basis of observation x. However, if we assume that ω is the chance element which caused the observation x and if $T_x(\omega) \subseteq H$, then, under this assumption ω, H must be necessarily true. So, it is surely of interest to examine the set of chance elements ω which imply H in this way,

$$u_x(H) = \{\omega \in v_x : T_x(\omega) \subseteq H\}.$$

We cannot know if the chance element which caused x belongs to this set, but we can compute the probability $P'(u_x(H))$ that this is the case. The larger this probability, the more hypothesis H becomes credible. Note that if $P'(u_x(H)) = 1$, then H must surely be true. On the other hand, $P'(u_x(H)) = 0$ does not yet mean that H is necessarily false (this will be discussed again below). We can see the $\omega \in u_x(H)$ as "arguments" in favor of H, and $P'(u_x(H))$ indicates the reliability of these arguments. Therefore,

$$sp_x(H) = P'(u_x(H))$$

is called the *degree of support* of H. Note carefully that the degree of support of a hypothesis is always relative to a functional model and an observation generated by it. The degree of support $sp_x(H)$ corresponds to the lower probabilities in the method of Dempster [**13, 16**] or to belief functions in Shafer's approach [**17**]. In fact, it will turn out that $sp_x(H)$ is formally a belief function in the sense of Dempster-Shafer theory of evidence. But in contrast to the work of Dempster and Shafer we obtain our degrees of support from a functional model and we interpret the degree of support as the reliability of inferences derived from this model and observations. So the semantics is quite different.

We mentioned above that a vanishing degree of support, $sp_x(H) = 0$ does not imply that H is necessarily false. This would only be the case, if $sp_x(H^c) = 1$, since then the complement H^c of H would be necessarily true. In this case all ω are arguments against H. In any case, it is interesting not only to consider

arguments in favor of H, but also arguments against H. The latter are simply arguments in favor of H^c. The larger the support for H^c, the less *plausible* H becomes. Therefore, the probability $P'(u_x(H^c))$ is called the *degree of doubt* of H. Or turned the other way round, we may say the less doubt we have in H, the more plausible the hypothesis is. Therefore we define the *degree of plausibility* of a hypothesis H by

$$pl_x(H) = 1 - sp_x(H^c).$$

This definition can be even more justified when we consider the chance elements which do not exclude H, that is, those that do not support H^c,

$$v_x(H) = u_x^c(H^c) = \{\omega \in v_x : T_x(H) \cap H \neq \emptyset\}. \tag{2.3}$$

This is the set of chance elements for which H remains possible, although not necessarily true. Note that

$$pl_x(H) = P'(v_x(H)).$$

Also note that $sp_x(H) \leq pl_x(H)$ for all subsets H of Θ. And $pl_x(H) = 0$ means that H is surely false. The degree of plausibility $pl_x(H)$ corresponds to the upper probabilities of Dempster and to the plausibility function of Shafer. But again our semantic is different: $pl_x(H)$ is the complement of the reliability of the deduction of H^c from the functional model and the observation.

So, using a functional model and an observation x generated by it, we may infer about the unknown parameter θ by determining "arguments", i.e. chance elements, which, if they would be the actual chance elements causing the observation, would allow to deduce H necessarily. This permits to compute degrees of support and plausibilities for any hypothesis and thus to judge their credibility. This may be the base for some decisions. But this comes only after the inference (see Section 5.1).

Let us illustrate this procedure of assumption-based reasoning with a few examples.

Example 2.4 (A simple coin). First we look at the model of a simple coin, example 2.1 above. Assume that $x = heads$ has been observed. Then both random elements remain possible, i.e. $v_{heads} = \{1, 2\}$. If we assume that face 1 turned up, i.e. $\omega = 1$, then $T_{heads}(1) = \{\theta_0, \theta_1\}$, i.e. under this assumption the experiment produces no information whatsoever about the unknown coin. If we assume however that face 2 turned up, i.e. $\omega = 2$, then $T_{heads}(2) = \{\theta_1\}$, i.e. we must necessarily conclude that the coin carries heads on both sides. Thus, we get $sp_{heads}(\theta_1) = 1/2$, whereas $sp_{heads}(\theta_0) = 0$ (for one-element sets we write $sp_{heads}(\{\theta_1\}) = sp_{heads}(\theta_1)$). There is no support for hypothesis θ_0. However, this hypothesis remains plausible to the degree $pl_{heads}(\theta_0) = 1 - sp_{heads}(\theta_1) = 1/2$. Nothing speaks against hypothesis θ_1, such that $pl_{heads}(\theta_1) = 1$.

If we observe $tails$, then, clearly, we must conclude that necessarily face 2 is turned up. So we have $v_{tails} = \{2\}$. It is clear in this case that hypothesis θ_0 must hold. This is reflected by the fact that $T_{tails}(2) = \{\theta_0\}$ and $sp_{tails}(\theta_0) = pl_{tails}(\theta_0) = 1$, whereas $sp_{tails}(\theta_1) = pl_{tails}(\theta_1) = 0$. △

EXAMPLE 2.5 (A sensor model). Next we examine the example of a test, according to example 2.3. Assume that the sensor signals an alarm, $x = a$. Note that both assumptions are still possible, $v_a = \{i, f\}$. If we assume the sensor is intact, then we must conclude necessarily that the hazardous material is present, $T_a(i) = \theta_1$, whereas, if we assume the sensor failed, then necessarily there is no hazardous material present, $T_a(f) = \theta_0$. From this we conclude

$$sp_a(\theta_1) = pl_a(\theta_1) = p, \quad sp_a(\theta_0) = pl_a(\theta_0) = 1 - p.$$

Similarly, if $x = s$ we obtain

$$sp_s(\theta_1) = pl_s(\theta_1) = 1 - p, \quad sp_s(\theta_0) = pl_s(\theta_0) = p.$$

In this example the sets $T_x(\omega)$ are all single element sets. In this particular situation, and only in this one, the degrees of support and of plausibility coincide, and also the degrees of support and plausibilities of complementary hypotheses sum up to one. △

EXAMPLE 2.6 (An urn model). Finally let us look at the urn model, example 2.2 above. But we propose to consider an alternative version of the model corresponding to an infinite population. That is, we take for Ω, Θ and X the unit interval $[0, 1]$ with uniformly distributed chance elements $\omega \in \Omega$. The functional model is

$$x = f(\theta, \omega) = \begin{cases} white & \text{if } \omega \leq \theta, \\ black & \text{otherwise.} \end{cases}$$

This model represents the limiting case of the finite urn model of example 2.2 if N is very large. This model violates the requirement of finite sets Ω, Θ and X, but the assumption-based analysis can be executed exactly as in the discrete case.

If we draw only one ball, the situation is very simple, but not very informative. Suppose the ball drawn is white, $x = white$. All random elements remain possible, $v_{white} = [0, 1]$. If we assume the chance element ω, then we conclude that θ must belong to $T_{white}(\omega) = \{\theta : \theta \geq \omega\}$. Similarly, if the ball drawn is black, then again all random elements remain possible (except $\omega = 0$, but this event has probability zero anyway) and $T_{black}(\omega) = \{\theta : \theta < \omega\}$.

Of more interest is the case when the experiment is executed n times with independent chance elements ω, corresponding to n draws of a ball with replacement. Let $\omega = (\omega_1, \ldots, \omega_n)$ be the corresponding random elements, the number of the balls drawn. Without loss of generality we may reorder and renumber the given sample such that the first x balls are white and the remaining $n - x$ are black. Then it becomes apparent that the first x random numbers ω_i must all be smaller than the last $n - x$ ones. So, we have

$$v_x = \{\omega : \max_{i=1,\ldots,x} \omega_i < \min_{j=x+1,\ldots,n} \omega_j\}. \tag{2.4}$$

In order to condition the probabilities we must compute the probability of v_x. We introduce the new random variables

$$Y_x = \max_{i=1,\ldots,x} \omega_i, \quad Z_{n-x} = \min_{j=x+1,\ldots,n} \omega_j.$$

The cumulative distribution functions of these two random variables are

$$P(Y_x \leq t) = t^x, t \in [0,1], \quad P(Z_{n-x} > s) = (1-s)^{n-x}, s \in [0,1],$$

and their respective density functions are xt^{x-1} and $(n-x)(1-s)^{n-x-1}$. The set v_x is now determined by the condition $Y_x < Z_{n-x}$. Its probability is

$$\begin{aligned} P(v_x) = P(Y_x < Z_{n-x}) &= \int_0^1 xt^{x-1}(1-t)^{n-x}dt \\ &= \frac{x\Gamma(x)\Gamma(n-x+1)}{\Gamma(n+1)} \end{aligned} \tag{2.5}$$

where $\Gamma(x)$ denotes the gamma function. Note that the inverse of $P(v_x)$ equals the binomial coefficient $\binom{n}{x}$.

If we fix a $\omega \in v_x$, then we can conclude that $T_x(\omega) = \{Y_x \leq \theta < Z_{n-x}\}$. This allows to compute degrees of support. For example, for hypotheses like $[a,b]$, where $0 \leq a < b \leq 1$, we obtain

$$\begin{aligned} sp_x(a \leq \theta \leq b) &= P'(T_x(\omega) \subseteq [a,b)) \\ &= \frac{\int_a^b xt^{x-1}((1-t)^{n-x} - (1-b)^{n-x})dt}{P(v_x)} \\ &= \int_a^b \frac{\Gamma(n+1)}{\Gamma(x)\Gamma(n-x+1)} t^{x-1}(1-t)^{n-x}dt \\ &- \binom{n}{x}(b^x - a^x)(1-b)^{n-x}. \end{aligned}$$

The integrand of the last integral is the density of the beta distribution with parameters x and $n-x+1$. If $be_{x,n-x+1}$ denotes the corresponding cumulative distribution function, we finally get

$$sp_x(a \leq \theta \leq b) = be_{x,n-x+1}(b) - be_{x,n-x+1}(a) - \binom{n}{x}(b^x - a^x)(1-b)^{n-x}.$$

Also of interest may be the plausibilities of singletons $\{\theta\}$,

$$pl_x(\theta) = P'(Y_x \leq \theta \leq Z_{n-x}) = \binom{n}{x}\theta^x(1-\theta)^{n-x}.$$

These plausibilities are proportional to the likelihood function associated with the experiment. Models that are similar to this urn model have been treated in [**13, 16**]. △

These examples show two things. First, the probability statements about the unknown parameter do not lead in general to a probability measure on the parameter space Θ as is suggested sometimes by fiducial probability. But still it

is possible to make probability statements about the unknown parameter with a very clear meaning: the degree of support of a hypothesis is the probability that the hypothesis can be logically deduced from the model and the data, i.e. the probability that sufficient assumptions hold true to derive the hypothesis. The degree of plausibility is the probability that the hypothesis cannot be refuted. Second, the likelihood function does not contain the full information about the parameter. In particular it makes no sense to normalize it to one and use it as a probability measure as has been sometimes proposed (for example in [**38**] part II).

An observation x related to a functional model together with the model itself represent a piece of information about the unknown parameter θ. If we take this point of view, we obtain an interesting structure. First we note that this information is *uncertain* in that it can be interpreted under different assumptions, and under each assumption we come to a certain conclusion about the unknown value. This resembles Shafer's random message model [**17**] for belief functions. The different possible assumptions are the possible values of ω in v_x. And if we select an assumption ω from v_x, then we conclude that the unknown value of θ must be in $T_x(\omega)$. T_x represents a multivalued mapping from v_x into the set Θ (as proposed by Dempster [**14**]). Finally, the possible assumptions in v_x have known probabilities $p'(\omega)$. Thus, finally, the information represented by the observation x can be summarized in a quadruple $\mathcal{H}_x = (v_x, p', T_x, \Theta)$. We call such a quadruple a *hint* [**7**]. Hints are of interest in themselves, since they allow to model any uncertain information, not only statistical data. Therefore, we introduce hints in general and discuss them in the following chapter.