Likelihood-Based Statistical Decisions

Marco E. G. V. Cattaneo
Seminar for Statistics, ETH Zürich
cattaneo@stat.math.ethz.ch

Abstract

In this paper, a nonadditive quantitative description of uncertain knowledge about statistical models is obtained by extending the likelihood function to sets and allowing the use of prior information. This description, which has the distinctive feature of not being calibrated, is called relative plausibility. It can be updated when new information is obtained, and it can be used for inference and decision making. As regards inference, the well-founded theory of likelihood-based statistical inference can be exploited, whereas decisions can be based on the minimax plausibility-weighted loss criterion. In the present paper, this decision criterion is introduced and some of its properties are studied, both from the conditional and from the repeated sampling point of view.

Keywords. Decision making, uncertainty, prior ignorance, minimax criterion, likelihood function, imprecise probabilities, nonadditive measure, completely maxitive measure, Shilkret integral, Choquet integral.

1 Introduction

In statistics, the likelihood function is considered as a measure, based on the data alone, of the relative plausibility of various models for those data. It plays a central role in many, if not most, modern procedures in all approaches to statistics, and some of the most appreciated inference methods are based directly on the likelihood function. It is therefore tempting to regard the likelihood function as a description of our uncertain knowledge about the models: a description that can be easily updated when new data are observed, and that can be used for inference and decision making. In this paper, a decision criterion based directly on the likelihood function (the minimax plausibility-weighted loss criterion) is proposed. It is defined in section 2, whereas in sections 3 and 4 some of its properties are studied: from the conditional and from the repeated sampling point of view, respectively.

2 Minimax Plausibility-Weighted Loss

2.1 Likelihood Function

Consider a set $\mathcal{P}$ of statistical models for an observed event A. That is, $\mathcal{P}$ is a set of probability measures[1] on a measurable space $(\Omega, \mathcal{A})$ such that $A \in \mathcal{A}$. The likelihood function on $\mathcal{P}$ based on the observation of A is the real-valued function $lik : P \mapsto P(A)$.

If we observe the realization x of a continuous random object X, we have $P\{X = x\} = 0$ for all $P \in \mathcal{P}$. But in reality, because of the finite precision of any observation, we only know that X lies in a neighborhood N of x (thus $lik(P) = P\{X \in N\}$). If f_P is a density of X under the model P, it can be useful to consider the approximation $P\{X \in N\} \approx \delta\, f_P(x)$. If this holds for all $P \in \mathcal{P}$, we obtain an approximate likelihood function, which is proportional to the function $P \mapsto f_P(x)$. This approximation can be very valuable, but sometimes leads to little problems, such as unbounded (approximate) likelihood functions. It should be emphasized that these problems are due to the above approximation and not to the likelihood function itself, which in particular is always bounded (since its values are probabilities).

The likelihood function lik is not calibrated, in the sense that only relative values are statistically relevant. The normed likelihood function $\overline{lik}$ (defined by $\overline{lik} \propto lik$ and $\sup_{P \in \mathcal{P}} \overline{lik}(P) = 1$) contains thus all the statistical information of lik, but its value $\overline{lik}(P)$ for a particular model P has no absolute meaning, even if it has often a simple asymptotic distribution (depending on the properties of $\mathcal{P}$). In fact, the general inference methods based directly on the likelihood

[1] Note that in statistics "model" is also used with a different meaning, to indicate a whole family of probability measures.

function (like the maximum likelihood estimator or the tests and confidence regions based on the likelihood ratio statistic) consider only relative values and may have to rely on large sample approximations.

In the literature on imprecise probabilities (for instance in [14], [17], [9], [3], [6] and [20]), there are many interesting proposals for obtaining possibilities, plausibilities, or upper probabilities (to be used then for inference or decision making) directly from the likelihood function. But they all give an absolute meaning to the values of lik or $\overline{lik}$ and are therefore at variance with the statistical meaning of the likelihood function. Exceptions are the likelihood-based inference method proposed in [19] and the axiomatic approach to decision making developed in [7].

2.2 Relative Plausibility

To underline the fact that the likelihood function is not calibrated, we introduce the concept of relative plausibility. A completely maxitive measure on $2^{\mathcal{P}}$ (see [16]) is a nonnegative, extended real-valued set function μ such that $\mu(\mathcal{H}) = \sup_{P \in \mathcal{H}} \mu\{P\}$ for all $\mathcal{H} \subseteq \mathcal{P}$, where (as in the rest of this paper) $\sup \varnothing = 0$. A relative plausibility rp on $\mathcal{P}$ is an equivalence class of completely maxitive measures on $2^{\mathcal{P}}$ with respect to the equivalence relation $\mu \sim \nu \Leftrightarrow \mu \propto \nu$. We can use rp to indicate one of its representatives only if the result is independent of the choice of the representative: for example, we can evaluate $rp\{P\}/rp\{P'\}$, but $rp\{P\}$ has no absolute meaning. By definition, a completely maxitive measure μ on $2^{\mathcal{P}}$ is uniquely determined by the function $\mu^{\downarrow} : P \mapsto \mu\{P\}$ on $\mathcal{P}$, and for each nonnegative, extended real-valued function f on $\mathcal{P}$ there is a unique relative plausibility rp on $\mathcal{P}$ such that $rp^{\downarrow} \propto f$ (rp is said to be generated by f).

A relative plausibility rp on $\mathcal{P}$ is interpreted as a description of our uncertain knowledge about the models in $\mathcal{P}$, and we are obviously interested in the one generated by the likelihood function. This is the extension of the relative likelihood to the subsets of $\mathcal{P}$ by means of the supremum. Such an extension can be conceptually debatable (although for $\mathcal{H}_0 \subset \mathcal{H}_1$ the ratio $rp(\mathcal{H}_0)/rp(\mathcal{H}_1)$ is the criterion of the likelihood ratio test of the hypothesis $P \in \mathcal{H}_0$ versus the hypothesis $P \in \mathcal{H}_1 \backslash \mathcal{H}_0$), but it is at least very useful from a notational standpoint.

If after having observed A our uncertain knowledge about the models in $\mathcal{P}$ is described by the relative plausibility rp, and we observe a new event $B \in \mathcal{A}$, we can easily update rp. The information obtained by the observation of B is encoded in the (conditional) likelihood function $lik' : P \mapsto P(B|A)$, and the updated relative plausibility is the one generated by $rp^{\downarrow}\, lik'$ (since $P(A \cap B) = P(A)\, P(B|A)$). In particular, if A and B are independent (under all models $P \in \mathcal{P}$), lik' is simply the likelihood function based on the observation of B (since $P(B|A) = P(B)$). It is thus natural to allow an expression of our uncertain knowledge about $\mathcal{P}$ prior to the observation of events in $\mathcal{A}$ in terms of a prior relative plausibility rp_{pre}, which is (informally) assumed to be based on information obtained from events independent of $\mathcal{A}$ (or equivalently, the models in $\mathcal{P}$ are assumed to be conditional on these events). After observing the first event $A \in \mathcal{A}$, we can combine the respective likelihood function lik with rp_{pre}, obtaining the relative plausibility generated by $rp_{pre}^{\downarrow}\, lik$. A (positive and finite) constant relative plausibility rp describes complete ignorance (all the models are equally plausible): in fact, using it as a prior is equivalent to using no prior (since $rp^{\downarrow}\, lik \propto lik$).

2.3 Decision Criterion

In a (statistical) decision situation, we have to choose an element of a set $\mathcal{D}$ of possible decisions, on the basis of a set $\mathcal{P}$ of considered models and of our uncertain knowledge about these models. A nonnegative loss function L on $\mathcal{P} \times \mathcal{D}$ is assumed to summarize all aspects of the possible decisions that should be considered in their evaluation:[2] $L(P, d)$ is the loss we would incur, according to the model P, by making the decision d. The quantitative comparison of two decisions is usually expressed in terms of the ratio of their losses (at least when these do not depend on $\mathcal{P}$). We shall therefore interpret the loss as a relative quantity: the loss functions $c\, L$ and L are equivalent, for all $c > 0$ (that is, the unit in which the loss is expressed is of no importance for the choice of a $d \in \mathcal{D}$). It is not necessary that $\inf_{d \in \mathcal{D}} L(P, d) = 0$ for all $P \in \mathcal{P}$, but it must be pointed out that if $c \neq 0$, the loss functions $L + c$ and L are not equivalent (since the ratio of the losses of two decisions depends on c), even if the most common decision criteria do not depend on c.

Consider the following situation: our uncertain knowledge about the models is described by a relative plausibility rp on $\mathcal{P}$ and we are faced with a decision problem described by a loss function L on $\mathcal{P} \times \mathcal{D}$. If we are completely ignorant about the considered models (that is, rp is constant), we can adopt the minimax criterion and choose a $d \in \mathcal{D}$ which minimizes $\sup_{P \in \mathcal{P}} L(P, d)$. A very intuitive way to extend this approach to the case in which we have some knowledge about the considered models (that is, rp is not constant), is to use $rp\{P\}$ as a weight for $L(P, d)$. This

[2] Note that the definitions of $\mathcal{D}$ and L are very general: $\mathcal{D}$ could for instance be a set of decision functions (in the sense of Wald) and L the respective risk.

leads to the MPL (Minimax Plausibility-weighted Loss) criterion: minimize

$$\sup_{P \in \mathcal{P}} rp\{P\} \, L(P,d). \tag{1}$$

A $d \in \mathcal{D}$ minimizing this quantity is called MPL decision; the set of MPL decisions does not depend either on the unit in which the loss is expressed or on the choice of the representative of rp. The MPL criterion is obviously parametrization invariant (since it is not based on a parametrization of $\mathcal{P}$), and is not concerned by the definition of rp for the subsets of $\mathcal{P}$ (other than the singletons). If there are several MPL decisions, their number should be reduced by using other criteria, in particular the inadmissible decisions should be discarded ($d \in \mathcal{D}$ is inadmissible[3] if there is a $d' \in \mathcal{D}$ such that $L(P,d) \geq L(P,d')$ for all $P \in \mathcal{P}$, and the inequality is strict for some $P \in \mathcal{P}$). However, usually the MPL decision is unique, and in this case it is certainly admissible.

2.4 Subadditive Integrals

The quantity (1) is proportional to the Shilkret integral of $L(\cdot,d)$ with respect to any representative μ of rp (see [16]). Completely maxitive measures are monotone and submodular, and for such measures the commonly used integral is the one of Choquet (see [5]). For completely maxitive measures on $2^{\mathcal{P}}$ and nonnegative functions on $\mathcal{P}$, the integrals of Shilkret and of Choquet have many common properties; in particular, they are subadditive:

$$\int (f + f') \, d\mu \leq \int f \, d\mu + \int f' \, d\mu.$$

On the contrary, the following properties are not shared by the two integrals. The Shilkret integral is maxitive:

$$\int \sup_{j \in J} f_j \, d\mu = \sup_{j \in J} \int f_j \, d\mu,$$

whereas the Choquet integral is additive for comonotonic functions: in particular, for all $c \in \mathbb{R}$ we have

$$\int (f + c) \, d\mu = \int f \, d\mu + c \, \mu(\mathcal{P}). \tag{2}$$

If we substitute the integral of Choquet for the one of Shilkret in the MPL criterion, we obtain the following criterion (let us call it MPL*): minimize the improper Riemann integral

$$\int_0^\infty rp\{P \in \mathcal{P} : L(P,d) > x\} \, dx.$$

A $d \in \mathcal{D}$ minimizing this quantity is called MPL* decision; the set of MPL* decisions does not depend either on the choice of the representative of rp or on the unit in which the loss is expressed. From the property (2) of the Choquet integral, it follows that if rp is finite, the MPL* criterion is invariant with respect to translations of the loss function (that is, the set of MPL* decisions for the problem described by the loss function $L + c$ does not depend on c). This invariance is very useful if the loss function is of the form $L = c - U$, where U is a real-valued utility function on $\mathcal{P} \times \mathcal{D}$ (that is, $U(P,d)$ represents the positive or negative gain in utility that we would obtain, according to the model P, by making the decision d; see for instance [2]). In fact, since utility functions are usually defined only up to positive affine transformations (that is, the utility functions $a\,U + b$ and U are equivalent, for all $a > 0$, $b \in \mathbb{R}$), the above invariance is almost necessary for a loss function of the form $L = c - U$. Unlike the Shilkret integral (which is defined only for nonnegative functions), the Choquet integral is defined also for real-valued functions, and the property (2) still holds. If the completely maxitive measure μ on $2^{\mathcal{P}}$ is finite and f is a real-valued function on $\mathcal{P}$, the Choquet integral satisfies $\int (-f) \, d\mu = -\int f \, d\overline{\mu}$, where the conjugate set function $\overline{\mu}$ on $2^{\mathcal{P}}$ is defined by $\overline{\mu}(\mathcal{H}) = \mu(\mathcal{P}) - \mu(\mathcal{P} \setminus \mathcal{H})$, for all $\mathcal{H} \subseteq \mathcal{P}$. Therefore, if the loss function is of the form $L = c - U$ and rp is finite, the MPL* criterion can be expressed as follows: maximize the Choquet integral $\int U(\cdot,d) \, d\overline{rp}$. Clearly, the set of decisions maximizing this integral is invariant with respect to positive affine transformations of U, and it does not depend on the choice of the representative of rp.

However, as already pointed out in subsection 2.3, in statistics the loss functions $L + c$ and L are generally not considered equivalent (if $c \neq 0$), since the loss is usually interpreted as a relative quantity. If we have a utility function U on $\mathcal{P} \times \mathcal{D}$, a loss function L whose relative meaning is invariant with respect to positive affine transformations of U can be defined as follows:

$$L(P,d) = \sup_{d' \in D} U(P,d') - U(P,d).$$

The loss obtained in this way from a utility is often called regret: $L(P,d)$ is the loss in utility that we suffer, according to the model P, for making the decision d instead of the optimal one (see [11]).

In the remainder of this paper we shall consider only the MPL criterion, but all the most important considered properties are satisfied also by the MPL* criterion. Moreover, we have seen that the MPL* criterion is invariant with respect to translations of the loss function (if rp is finite): though not necessary for a loss function with a relative meaning, this invari-

[3] This definition generalizes the classical one, which is limited to the case mentioned in footnote [2].

ance can be useful. But a basic feature of the MPL criterion is lost by the MPL* criterion: its extreme simplicity.

2.5 Pseudo Likelihood

Let $\mathcal{P}$ be a set of models and let lik be the likelihood function on $\mathcal{P}$ based on the observation of some event. Consider a mapping $T : \mathcal{P} \to \mathcal{T}$, which assigns to each model the respective value of the parameter of interest (that is, $T(P)$ is the only aspect of P in which we are interested). In particular, if T is bijective, T is called a parametrization of $\mathcal{P}$, and $lik \circ T^{-1}$ is simply called likelihood function on $\mathcal{T}$. Obviously, for all purposes of inference or decision making we can use the likelihood function on $\mathcal{T}$ instead of the one on $\mathcal{P}$. In general, for any mapping T, a pseudo likelihood function $\widetilde{lik}$ on $\mathcal{T}$ is a function which, to some extent at least, can be used as if T were a parametrization of $\mathcal{P}$ and $\widetilde{lik}$ were the likelihood function on $\mathcal{T}$. The profile likelihood function based on the mapping T is the function lik_T on $\mathcal{T}$ defined by $lik_T(t) = \sup_{P \in T^{-1}\{t\}} lik(P)$. The profile likelihood functions are the simplest and most important kind of pseudo likelihood functions.

Consider the following situation: we have a relative plausibility rp on $\mathcal{P}$ and we are faced with a decision problem described by a loss function L which depends on the model through the parameter of interest only. That is, $L(P, d) = L'[T(P), d]$ for some nonnegative function L' on $\mathcal{T} \times \mathcal{D}$. In this case, the MPL criterion can be expressed as follows: minimize

$$\sup_{P \in \mathcal{P}} rp\{P\}\, L(P, d) = \sup_{t \in \mathcal{T}} rp\{T = t\}\, L'(t, d),$$

where, as usual, $\{T = t\}$ simply means $T^{-1}\{t\}$. In particular, if rp is generated by the likelihood function lik, the MPL criterion is simply the following: minimize

$$\sup_{P \in \mathcal{P}} lik(P)\, L(P, d) = \sup_{t \in \mathcal{T}} lik_T(t)\, L'(t, d).$$

That is, the MPL criterion automatically considers the profile likelihood function of the parameter of interest.

Although the use of the profile likelihood function usually leads to reasonable results, better results can sometimes be achieved by using other pseudo likelihood functions. In the literature on likelihood-based statistical inference, many alternative pseudo likelihood functions (such as conditional, marginal, modified profile, partial, integrated or estimated likelihood functions) have been proposed for different situations (see for instance [13] and [10]). Obviously, if some pseudo likelihood function $\widetilde{lik}$ on $\mathcal{T}$ is expected to give better results than lik_T, it should be used. This leads to a pseudo MPL criterion: minimize

$$\sup_{t \in \mathcal{T}} \widetilde{lik}(t)\, L'(t, d).$$

3 Conditional Point of View

Consider the following situation: we have a set $\mathcal{P}$ of statistical models for an observed event A, we can have some prior uncertain knowledge about the models in $\mathcal{P}$, and we are faced with a decision problem described by a loss function L on $\mathcal{P} \times \mathcal{D}$. The whole decision situation can be conditional on the observation A, in the following sense: the alternative events which could have been observed (that is, a partition of the event A^c) can be undefined, and for each one of these alternative events the possible decisions and the loss can be undefined (it can be impossible to imagine which decision problem we would have faced if we had observed an alternative event). To be applicable in such a situation, a decision criterion must obviously avoid the consideration of the possible alternative events and the respective decision problems: a criterion with this property is called conditional. Conditional criteria are interesting also if the decision situation is not conditional on the observation A (that is, we can define the possible alternative events and the respective decision problems), because we may want our decision to be based only on the event which was actually observed. In this sense, conditional criteria are based on a post-data evaluation of the possible decisions, as opposed to a pre-data evaluation considering all the possible observations.

A decision criterion satisfies the strong likelihood principle if it is conditional and it depends on the observed event through the respective relative likelihood function only. It can be argued that a conditional criterion which is at the same time general and reasonable must satisfy this principle. Anyway, a criterion which satisfies it must use in some way the prior information about the models, the relative likelihood function based on the observation, and the loss function. The most elegant and clear way of doing this is to compare the different decisions on the basis of the loss function and of some quantitative description of the uncertain knowledge about the models, a description that can be updated through the relative (conditional) likelihood functions based on the observations.

3.1 Quantitative Descriptions of Uncertain Knowledge about Statistical Models

The most important statistical theory based on a quantitative description of the uncertain knowledge about the models is obviously the Bayesian one. In this theory, the uncertain knowledge is described by a probability measure on $\mathcal{P}$. This measure can be updated through the relative (conditional) likelihood functions based on the observations, and it can be used straightforwardly for inference and decision making. The Bayesian theory has many important properties, whereas its main problem is the need of a prior probability measure on $\mathcal{P}$.

Many generalizations of the Bayesian theory allowing an imprecise prior have been proposed (see for instance [2] and [1]). Most of them are formally equivalent with the choice of some set of prior probability measures on $\mathcal{P}$ (although the interpretations of this set can be very different). These measures can be individually updated through the relative (conditional) likelihood functions based on the observations, and the set of probability measures can be used in some way (depending on the interpretation; see for instance [12]) for inference and decision making. Although these approaches allow some imprecision in the prior, they have problems with the representation of ignorance. In particular, if we choose the set of all probability measures on $\mathcal{P}$ as our imprecise prior, the inferences will usually remain vacuous independently of the amount of statistical information obtained from the observations. In every situation, some compromise between the ignorance in the prior and the power of the inferences must be reached (see for example [19]).

In this paper, the direct use of the relative likelihood (or its extension to sets: the relative plausibility) as a description of the uncertain knowledge about the models is proposed. This description can obviously be updated through the relative (conditional) likelihood functions based on the observations, and it can be used for inference and decision making. The likelihood-based statistical inference is a well-founded theory, whose conclusions are in general weaker than those based on probabilistic assumptions (but which do not need these assumptions), because the likelihood function is not calibrated (see [10]). A calibration based on a repeated sampling interpretation is usually possible, and can be very simple if a large sample approximation applies (see [13]). As regards decision making, in subsection 2.3 the MPL criterion has been introduced. This criterion is based only on the relative likelihood: in fact, no absolute meaning is given to the quantity (1), which is used only in a relative way, to compare the decisions $d \in \mathcal{D}$.

3.2 Prior Uncertain Knowledge

If we are able to define a prior probability measure π on $\mathcal{P}$, and we are ready to give to this measure the same status of the elements of $\mathcal{P}$, we should do it: what we obtain is a single probability measure P_π on $\mathcal{P} \times \Omega$. If (after having observed the event $A \in \mathcal{A}$) we are faced with a decision problem described by a loss function L on $\mathcal{P} \times \mathcal{D}$, we can define the new loss function L' on $\{P_\pi\} \times \mathcal{D}$ by $L'(P_\pi, d) = E_{P_\pi}[L(P, d) | \mathcal{P} \times A]$. The MPL criterion applied to the decision problem described by L' corresponds to the (conditional) Bayesian criterion: minimize the posterior expected loss $L'(P_\pi, d)$.

If we are not able to precisely define the prior probability measure π, but we maintain that it belongs to a particular set Γ of probability measures on $\mathcal{P}$, we obtain a new set of models $\mathcal{P}' = \{P_\pi : \pi \in \Gamma\}$, and we can define L' on $\mathcal{P}' \times \mathcal{D}$ as above. If we get no information from the observation of A (that is, the respective likelihood function on $\mathcal{P}$ is constant), the MPL criterion applied to the decision problem described by L' corresponds to the (conditional) Γ-minimax criterion: minimize $\sup_{\pi \in \Gamma} L'(P_\pi, d)$. But if we get some information from the observation of A, the MPL criterion applied to the decision problem described by L' is the following: minimize $\sup_{\pi \in \Gamma} E_\pi[P(A)] \, L'(P_\pi, d)$. That is, in this case the MPL criterion uses the (second-order) likelihood $E_\pi[P(A)]$ as a weight for the posterior expected loss with respect to π. The (second-order) likelihood function $lik' : P_\pi \mapsto P_\pi(\mathcal{P} \times A) = E_\pi[P(A)]$ on $\mathcal{P}'$ allows non-vacuous inferences, even if Γ is the set of all probability measures on $\mathcal{P}$ (that is, it allows us to get out of the state of complete ignorance). Since the relative plausibility on $\mathcal{P}'$ generated by the likelihood function lik' is proportional to a (second-order) possibility measure, we can consider this description of our uncertain knowledge about the models in $\mathcal{P}$ as a non-calibrated possibilistic hierarchical model (see [4]).

If we are not ready to give to our prior uncertain knowledge about the models in $\mathcal{P}$ a fully probabilistic status, we can describe it by means of a prior relative plausibility. This can be based on analogies with past experience in a very natural way: in fact, what we can observe about models are relative likelihoods, not probabilities.

Relative plausibility seems to be more intuitive than probability: for example, the intuitive idea that a diffuse prior represents ignorance is wrong for probability, whereas it is correct for relative plausibility. In fact, as noted at the end of subsection 2.2, a constant relative plausibility describes complete ignorance. Partial ignorance can also be easily described: consider for instance that we have a mapping $T : \mathcal{P} \to \mathcal{T}$ and that the function f on $\mathcal{T}$ describes the prior relative likelihood of the different values of the parameter of interest $T(P)$. If we are otherwise ignorant about $\mathcal{P}$, we can use the prior relative plausibility rp generated by $f \circ T$, since the plausibility ratio $rp\{P\}/rp\{P'\}$ is then simply $f[T(P)]/f[T(P')]$ (for all $P, P' \in \mathcal{P}$). If we update rp through the likelihood function lik (based on the observation of A), the plausibility ratio $rp\{T = t\}/rp\{T = t'\} = f(t)/f(t')$ is simply multiplied by the factor $lik_T(t)/lik_T(t')$ (for all $t, t' \in \mathcal{T}$), where lik_T is the profile likelihood function based on T (defined in subsection 2.5). This is a very intuitive result: for instance, the case $|T| = 2$ can be described as follows. Let $\mathcal{H}_0$ and $\mathcal{H}_1$ be two disjoint nonempty sets such that $\mathcal{H}_0 \cup \mathcal{H}_1 = \mathcal{P}$. If our prior uncertain knowledge about the models is limited to the plausibility ratio $rp(\mathcal{H}_0)/rp(\mathcal{H}_1)$, after the observation of A we have

$$\frac{rp'(\mathcal{H}_0)}{rp'(\mathcal{H}_1)} = \frac{\sup_{P \in \mathcal{H}_0} lik(P)}{\sup_{P \in \mathcal{H}_1} lik(P)} \, \frac{rp(\mathcal{H}_0)}{rp(\mathcal{H}_1)},$$

where rp' is the updated relative plausibility (that is, the one generated by $rp^{\downarrow}\, lik$). Compare this result with the so-called Lindley's paradox (see [15]).

It is important to note also that two prior relative plausibilities rp and rp' on $\mathcal{P}$ which are assumed to be based on independent observations can be easily combined: the resulting relative plausibility is the one generated by $rp^{\downarrow}\, rp'^{\downarrow}$.

3.3 Nonadditivity

The relative plausibility can be considered as an imprecise probability, in the sense that it is a nonadditive quantitative description of uncertain knowledge. Its principal distinctive feature is to be only a relative measure: in fact, a relative plausibility rp on $\mathcal{P}$ is mathematically defined as an equivalence class of proportional (completely maxitive) measures on $2^{\mathcal{P}}$. Since a completely maxitive measure μ on $2^{\mathcal{P}}$ such that $\mu(\mathcal{P}) \leq 1$ is a possibility measure on $\mathcal{P}$, the name "relative possibility" would better describe the mathematical properties of rp, but "plausibility" better describes the meaning attached to it.

The nonadditivity of rp implies in particular that it is impossible to define an additive integral with respect to (a representative of) rp, since $\int I_{\mathcal{H}}\, d\mu = \mu(\mathcal{H})$ is required. The nonadditivity of the integral has inconvenient consequences: for example, consider two decision problems described by the loss functions L_1 on $\mathcal{P} \times \mathcal{D}_1$ and L_2 on $\mathcal{P} \times \mathcal{D}_2$, respectively. We can combine them in the single decision problem described by the loss function L on $\mathcal{P} \times (\mathcal{D}_1 \times \mathcal{D}_2)$ defined by $L[P, (d_1, d_2)] = L_1(P, d_1) + L_2(P, d_2)$. Because of the nonadditivity of the integral, if $\widetilde{d_1}$ and $\widetilde{d_2}$ are MPL decisions for the two problems considered separately, $(\widetilde{d_1}, \widetilde{d_2})$ needs not to be a MPL decision for the compound problem. In fact, since the Shilkret integral is maxitive, $(\widetilde{d_1}, \widetilde{d_2})$ is a MPL decision for the compound problem if L is defined by $L[P, (d_1, d_2)] = \max\{L_1(P, d_1), L_2(P, d_2)\}$.

3.4 Robustness

The MPL decisions are invariant with respect to the consideration of additional statistical models with sufficiently small relative plausibilities. More precisely, if $\mathcal{P}' \subseteq \mathcal{P}$, rp is a relative plausibility on $\mathcal{P}$, L is a loss function on $\mathcal{P} \times \mathcal{D}$, d is a MPL decision for the problem described by the restriction of L to $\mathcal{P}' \times \mathcal{D}$ (with as relative plausibility on $\mathcal{P}'$ the restriction of rp), and $rp\{P\} \leq c/L(P, d)$ for all $P \in \mathcal{P} \setminus \mathcal{P}'$ (where $c = \sup_{P' \in \mathcal{P}'} rp\{P'\}\, L(P', d)$), then d is also a MPL decision for the problem described by L. Therefore, instead of considering a set $\mathcal{P}'$ of models, we can extend it to a broader set $\mathcal{P}$ (or even to the whole set of all probability measures on $(\Omega, \mathcal{A})$) and use a prior relative plausibility on $\mathcal{P}$ to describe the prior likelihood of the different models: the models in $\mathcal{P} \setminus \mathcal{P}'$ can influence the MPL decisions only if their (updated) relative plausibilities are large enough.

Let $\mathcal{IP}$ be a set of imprecise models, in the sense that every $IP \in \mathcal{IP}$ is a set of probability measures on $(\Omega, \mathcal{A})$. We can consider this situation by defining $\mathcal{P} = \cup \mathcal{IP}$ and $T : \mathcal{P} \to \mathcal{IP}$ such that $T^{-1}\{IP\} = IP$ (some technicalities are needed if the sets IP overlap). A prior relative plausibility on $\mathcal{P}$ based on a prior relative likelihood function on $\mathcal{IP}$ can be defined as described in subsection 3.2, and if lik is the likelihood function on $\mathcal{P}$ based on the observation of the event $A \in \mathcal{A}$, then $lik_T(IP) = \sup_{P \in IP} P(A)$. That is, using imprecise models simply means considering the pseudo likelihood function on $\mathcal{IP}$ defined by the upper probabilities. In particular, if $\mathcal{P}$ is a set of models, replacing every $P \in \mathcal{P}$ by the corresponding ε-contamination class (that is, the set of all models of the form $(1 - \varepsilon)\, P + \varepsilon\, P'$, for all probability measures P' on $(\Omega, \mathcal{A})$) simply means considering the pseudo likelihood function $(1 - \varepsilon)\, lik + \varepsilon$ on $\mathcal{P}$ instead of the likelihood function lik (this can be interpreted as a discounting of the information encoded in lik; see [14]).

4 Repeated Sampling Point of View

Consider the following situation: we have a set $\mathcal{P}$ of statistical models for a random object $X : \Omega \to \mathcal{X}$, we can have some prior uncertain knowledge about the models in $\mathcal{P}$, and for each possible realization x of X we have a decision problem described by a loss function L_x on $\mathcal{P} \times \mathcal{D}_x$. Since this decision situation is not conditional on the observation of a particular event $\{X = x\}$, we can consider a pre-data evaluation of the decisions, in the following sense. A decision function δ on $\mathcal{X}$ assigns to each possible realization x of X a decision $\delta(x) \in \mathcal{D}_x$ for the problem described by L_x. Under each model $P \in \mathcal{P}$, the performance of δ is described by the function $x \mapsto L_x[P, \delta(x)]$ on $\mathcal{X}$; if this function is measurable, the (pre-data) evaluation of δ can be based on the random variable $L_X[P, \delta(X)]$. If the different losses L_x are expressed in the same unit, we can consider $L_X[P, \delta(X)]$ as the random loss of δ, and in this case the simplest way to compare different decision functions is to reduce the random loss to a single (representative) value. The expected value of the loss (called risk) is generally used, but sometimes other aspects of the random loss (such as quantiles) are considered. Anyway, for each model $P \in \mathcal{P}$ and each decision function $\delta \in \mathcal{D}$ (the set of the possible decision functions), we have a representative value $L(P, \delta)$ of the random loss. Considered in this way, the choice of a decision function corresponds to the (pre-data) decision problem described by the loss function L on $\mathcal{P} \times \mathcal{D}$. If $L(P, \delta)$ is the risk of δ under the model P and we have no prior information about the models in $\mathcal{P}$, the MPL criterion applied to this problem corresponds to the usual minimax risk criterion; but if we have a prior relative plausibility rp on $\mathcal{P}$, the MPL criterion uses $rp\{P\}$ as a weight for the risk with respect to P.

Since the MPL decision criterion is conditional, we can also apply it after the observation of the realization x of X, to the problem described by L_x (using the updated relative plausibility). A conditional MPL decision function δ is a function on $\mathcal{X}$ such that (for all $x \in \mathcal{X}$) $\delta(x)$ is a MPL decision for the conditional problem described by L_x. In general, δ is not a MPL decision for the problem described by L: this implies that the choice between pre-data and post-data evaluation is important.[4] Since the pre-data decision problem is usually much more difficult than the conditional problems, it is interesting to study the properties of the conditional MPL decision functions from the repeated sampling point of view, in particular if L_x and $\mathcal{D}_x$ do not depend on x. In this case, the observation of the realizations of one or more random objects can be interpreted as a way to acquire information useful for making a better choice in a particular decision problem.

4.1 Equivariance

Consider the following situation: we have a set $\mathcal{P}$ of statistical models for a random object $X : \Omega \to \mathcal{X}$, and we are faced with a decision problem described by a loss function L on $\mathcal{P} \times \mathcal{D}$. Let δ be a conditional MPL decision function on $\mathcal{X}$ (obtained without using a prior relative plausibility). If the decision problem is invariant (under a group of transformations of $\mathcal{X}$), then δ is equivariant (see for instance [2], where this property is called invariance). More precisely, δ is equivariant if it is unique; if it is not unique, the equivariance property still holds, but for the sets of MPL decisions. If we use a prior relative plausibility rp on $\mathcal{P}$, these results hold only if rp satisfies an invariance property. The proofs are straightforward, but the introduction of the elements of an invariant problem would need too much space.

When the decision problem is invariant, the equivariance is generally considered as a very important property for a decision function, assuring that this presents symmetries corresponding to those of the problem. The equivariance is usually a restriction imposed on the set of the possible decision functions, after consideration of the symmetries of the problem. The conditional MPL decision criterion guarantees the equivariance of the corresponding decision function, without need of considering the symmetries (and only if these are not invalidated by asymmetric prior information).

4.2 Asymptotic Optimality

If we have a decision problem described by a loss function L on $\mathcal{P} \times \mathcal{D}$, under the model P an optimal decision $d \in \mathcal{D}$ is obviously one that minimizes the loss $L(P, d)$. Consider two different models $P, P' \in \mathcal{P}$; if we observe a sequence of realizations of random objects, we can expect that, according to the model P, the plausibility ratio of P and P' tends to infinity as the number of observations tends to infinity. Since this is valid for all P' different from P, and since the relative plausibility is used in the MPL criterion as a weight for the loss, we can expect that, under the model P, the loss $L(P, d)$ tends to have a decisive importance in the choice of d. Thus we can expect that, according to any model in $\mathcal{P}$, the conditional MPL decisions tend to be optimal as the number of observations tends to infinity: this property can be called asymptotic optimality. In fact, this property can be easily shown for the case of a finite set of models $\mathcal{P}$ (and even in the strong sense of almost sure convergence), whereas for the case of an infinite set of models some regularity conditions on the models and on the loss function are needed (and stronger conditions are needed for assuring almost sure convergence).

The asymptotic optimality can be considered as a minimal requirement for a decision criterion based on statistical information. It is important to note that the asymptotic optimality is not affected by the use of a prior relative plausibility, as long as this satisfies

[4] The property of equivalence between pre-data and post-data evaluation is satisfied by the Bayesian expected loss criterion (but only if L is defined as the risk); see [1].

some regularity conditions (such as being positive).

4.3 Pre-Data Evaluation

The MPL criterion satisfies the strong likelihood principle. A decision criterion satisfying this principle considers the observed events through the respective relative likelihood functions only. Since the same likelihood function can be obtained from very different problems, the decision function obtained by the post-data application of a criterion satisfying the strong likelihood principle can have arbitrarily bad pre-data properties in some particular problems, if the criterion does not use prior information. In practice, the same is true even if it uses prior information, as long as this has not been chosen precisely to assure good pre-data properties. Besides the use of a prior relative plausibility, the conditional MPL criterion has another possibility for obtaining better pre-data properties for the corresponding decision functions: the use of pseudo likelihood functions. In fact, the pseudo MPL criterion does not in general satisfy the strong likelihood principle.

On the other side, a decision criterion satisfying the strong likelihood principle has the fundamental property of being independent of the choice of a sufficient statistic for the data. Moreover, since such a criterion is conditional, it can be applied post-data, avoiding the important problems related to the ancillary statistics (see for instance [8]), and drastically reducing the complexity of the decision problem. In fact, a conditional MPL decision can be found (or at least numerically approximated) even in difficult problems, in which it is practically impossible to find a decision function satisfying some kind of pre-data optimality.

4.4 Examples

As examples of conditional MPL decision functions, we can consider the ones obtained in some of the simplest and most important estimation problems: so we can compare the risk of these decision functions with the risk of the usual estimators. These comparisons make sense only if we start with complete ignorance and we consider the usual loss functions (in general, the MPL decisions depend on the choice of the loss function).

4.4.1 Binomial

The estimation of the probability parameter p of the binomial distribution $\mathcal{B}(n,p)$ is certainly one of the most important estimation problems for discrete distributions. For a fixed n, let $\mathcal{P} = \{P_p : p \in [0,1]\}$, with $X \sim_{P_p} \mathcal{B}(n,p)$. We can consider, as usual, the squared error loss $L(P_p, d) = (d-p)^2$, where $d \in \mathcal{D} = [0,1]$. For $n = 1$, the conditional MPL decision function δ_{MPL} corresponds to the minimax risk decision function δ_{MR} (that is, $\delta_{MPL}(0) = \frac{1}{4}$ and $\delta_{MPL}(1) = \frac{3}{4}$). But for $n > 1$, δ_{MPL} leaves δ_{MR} and, as n increases, it tends rapidly toward the maximum likelihood decision function δ_{ML} (which is the usual estimator: $\delta_{ML}(x) = \frac{x}{n}$). This is a good behavior, since usually δ_{MR} is preferred (on the basis of the squared error risk) for small values of n, whereas δ_{ML} is preferred for large (and even for moderate) values of n.

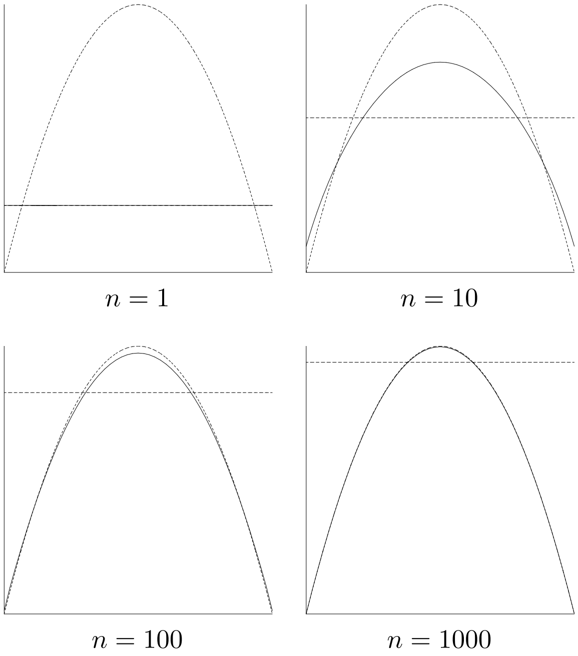

Figure 1: Risk functions of δ_{MPL}, δ_{MR} and δ_{ML}.

The four graphs of Figure 1 show the risk functions $p \mapsto E_{P_p}[(\delta(X) - p)^2]$ of $\delta = \delta_{MPL}$ (the solid lines), $\delta = \delta_{MR}$ (the constant dashed lines) and $\delta = \delta_{ML}$ (the parabolic dashed lines), for different values of n (in each graph, the abscissa axis is $[0,1]$ and the ordinate axis is $[0, \frac{1}{4n}]$). For $n = 1$, the risk function of δ_{MPL} corresponds to the one of δ_{MR}, whereas for $n = 1000$, it (nearly) corresponds to the one of δ_{ML}.

4.4.2 Normal

The estimation of the parameters μ and σ^2 of the normal distribution $\mathcal{N}(\mu, \sigma^2)$ on the basis of a sample $x = (x_1, \ldots, x_n)$ is certainly one of the most important estimation problems for continuous distributions. Let $\mathcal{P} = \{P_{\mu,\sigma} : \mu \in \mathbb{R},\ \sigma > 0\}$, with $X_1, \ldots, X_n \overset{iid}{\sim}_{P_{\mu,\sigma}} \mathcal{N}(\mu, \sigma^2)$. We can consider, as usual, the squared error losses, and in particular (since $\mathcal{P}$ is a location-scale family of models) their invariant versions: $L(P_{\mu,\sigma}, d) = \frac{(d-\mu)^2}{\sigma^2}$ (where $d \in \mathcal{D} = \mathbb{R}$) for

n	2	5	10	20	50
δ_{MPL}	0.996	0.995	0.996	0.998	0.999
δ_{MVU}	0.333	0.667	0.818	0.905	0.961
δ_{ML}	0.889	0.926	0.957	0.977	0.990

Table 1: Relative efficiencies with respect to δ_{MRE}.

the estimation of μ, and $L(P_{\mu,\sigma}, d) = \frac{(d-\sigma^2)^2}{\sigma^4}$ (where $d \in \mathcal{D} = (0, \infty)$) for the estimation of σ^2. Moreover, as usual, we can base our conclusions on the approximate likelihood function obtained from the density of $X = (X_1, \ldots, X_n)$.

The conditional MPL decision function for the estimation of μ corresponds to the usual (and undisputed) estimator, which assigns to each x the arithmetic mean $\overline{x} = \frac{x_1+\cdots+x_n}{n}$. For $n \geq 2$, the conditional MPL decision function for the estimation of σ^2 is $\delta_{MPL} = \frac{s^2}{n+c_n}$, where $c_n \approx 1.3$ (c_n varies slightly with n) and s^2 is the function which assigns to each x the sum of squares $\sum_{j=1}^{n}(x_j - \overline{x})^2$. The usual estimator of σ^2 is the (uniform) minimum variance unbiased decision function $\delta_{MVU} = \frac{s^2}{n-1}$, whereas the maximum likelihood decision function is $\delta_{ML} = \frac{s^2}{n}$. Since δ_{MPL}, δ_{MVU} and δ_{ML} are location-scale equivariant, we can compare their constant risks with the one of the minimum risk (location-scale) equivariant decision function $\delta_{MRE} = \frac{s^2}{n+1}$. The ratio of the constant risks of δ_{MRE} and δ is sometimes called the relative efficiency of δ with respect to δ_{MRE}, and is given in Table 1 for $\delta = \delta_{MPL}$, $\delta = \delta_{MVU}$ and $\delta = \delta_{ML}$, for different values of n. We see that δ_{MPL} is almost optimal (among the equivariant decision functions) even for small values of n; but as n increases, the optimality is approached also by δ_{MVU} and δ_{ML}.

4.4.3 Asymmetric Error

In all the estimation problems considered above, we have used as loss $L(P, d)$ the (weighted) squared error: this error is symmetric, in the sense that the function $L(P, \cdot)$ is symmetric about the estimated parameter, for all $P \in \mathcal{P}$. In an estimation problem with symmetric error, the conditional MPL decision function δ_{MPL} usually tends rapidly toward the maximum likelihood decision function δ_{ML}, as the number of observations increases. This is good, since δ_{ML} is often asymptotically efficient, and δ_{MPL} usually shares this property. But if the error is asymmetric, the asymptotic efficiency can be a drawback for a decision function, and in this case δ_{MPL} often tends toward δ_{ML} too slowly to be asymptotically efficient. Estimation problems with asymmetric error are not very common in the statistical literature, but we can consider for instance the two examples studied in [18]: these can be regarded as the first published examples of application of the general theory of statistical decisions.

In the first example, we consider the estimation of the mean μ of the normal distribution $\mathcal{N}(\mu, 1)$ with variance 1, on the basis of a sample $x = (x_1, \ldots, x_n)$. Let $\mathcal{P} = \{P_\mu : \mu \in \mathbb{R}\}$, with $X_1, \ldots, X_n \overset{iid}{\sim}_{P_\mu} \mathcal{N}(\mu, 1)$. As loss we use the asymmetric error

$$L(P_\mu, d) = \begin{cases} 2\,(d-\mu) & \text{if } d \geq \mu \\ \mu - d & \text{if } d \leq \mu \end{cases}$$

(where $d \in \mathcal{D} = \mathbb{R}$), and, as usual, we base our conclusions on the approximate likelihood function obtained from the density of $X = (X_1, \ldots, X_n)$. The considered asymmetric error favors the underestimation of μ, and in fact $\delta_{MPL}(x) = \overline{x} - \frac{c}{\sqrt{n}}$, with $c \approx 0.345$. Since the maximum likelihood estimator does not consider the error, we have $\delta_{ML}(x) = \overline{x}$; whereas the minimax risk decision function δ_{MR} has the same form of δ_{MPL}, with $c \approx 0.431$. Since the estimation problem is location invariant, and δ_{MPL}, δ_{ML} and δ_{MR} are location equivariant, we can compare their constant risks. The relative efficiency of δ_{MPL} with respect to δ_{MR} is 0.996, whereas the one of δ_{ML} is 0.911. These relative efficiencies are independent of n, and in fact δ_{MPL} and δ_{MR} (unlike δ_{ML}) are not asymptotic efficient estimators of μ.

In the second example, we consider the estimation of the mean μ of the uniform distribution $\mathcal{U}(\mu - \frac{1}{2}, \mu + \frac{1}{2})$, on the basis of a sample $(x_1, \ldots, x_n)$. We use the same asymmetric error as in the previous example, and we get similar results, with the difference that this time $\delta_{MPL} = \delta_{MR}$ (whereas δ_{ML} is not uniquely defined). Since the estimation problem is location invariant, δ_{MPL} is the minimum risk (location) equivariant decision function.

5 Conclusion

In the present paper, the minimax plausibility-weighted loss criterion has been introduced: this allows decisions to be based directly on the likelihood function. Therefore, by extending the likelihood function to sets and allowing the use of prior information, we obtain a full-fledged quantitative description of the uncertain knowledge about the models: the relative plausibility. This can be considered as a non-calibrated imprecise probability (or more precisely, as a non-calibrated possibility measure) on the considered set of models. The conclusions based on the relative plausibility are in general weaker than those based on a probabilistic description of the uncertain knowledge about the models; but the relative plausibility is based on weaker assumptions, it is simpler and more intuitive.

The minimax plausibility-weighted loss criterion is a simple, intuitive and completely general decision criterion. It has many important properties, such as parametrization invariance, independence of the choice of a sufficient statistic for the data, conditionality, possibility of using prior information, but also of starting with complete ignorance, some kind of robustness with respect to the consideration of additional models, possibility of using pseudo likelihood functions and of considering imprecise models. The conditional application of the criterion allows decisions even in difficult problems, and the obtained decision functions are equivariant (if the problem is invariant) and asymptotic optimal (if some regularity conditions are satisfied), but they can have bad pre-data properties in particular problems (as any decision function obtained by the conditional application of a criterion satisfying the strong likelihood principle). The consideration of many examples seems to suggest that the minimax plausibility-weighted loss criterion leads in general to reasonable decisions. Some of these examples (for instance the estimation of variance components in mixed effects models), as well as the exact statements and proofs of the properties considered in this paper, will be published in my PhD thesis, which will be submitted soon.

References

[1] T. Augustin. *On the Suboptimality of the Generalized Bayes Rule and Robust Bayesian Procedures from the Decision Theoretic Point of View: A Cautionary Note on Updating Imprecise Priors.* ISIPTA '03 (eds. Bernard, Seidenfeld and Zaffalon), 31–45, Carleton Scientific, 2003.

[2] J. Berger, *Statistical Decision Theory and Bayesian Analysis* (2nd edition). Springer, 1985.

[3] Y. Y. Chen. *Statistical Inference Based on the Possibility and Belief Measures.* Trans. Am. Math. Soc. 347:1855–1863, 1995.

[4] G. de Cooman and P. Walley. *A Possibilistic Hierarchical Model for Behaviour under Uncertainty.* Theory Decis. 52:327–374, 2002.

[5] D. Denneberg. *Non-Additive Measure and Integral.* Kluwer, 1994.

[6] D. Dubois, S. Moral and H. Prade. *A Semantics for Possibility Theory Based on Likelihoods.* J. Math. Anal. Appl. 205:359–380, 1997.

[7] P. H. Giang and P. P. Shenoy. *Decision Making on the Sole Basis of Statistical Likelihood.* Artif. Intell. (in press), 2005.

[8] C. Goutis and G. Casella. *Frequentist Post-Data Inference.* Int. Stat. Rev. 63:325–344, 1995.

[9] S. Moral, L. M. de Campos. *Updating Uncertain Information.* Uncertainty in Knowledge Bases (eds. Bouchon-Meunier, Yager and Zadeh), 58–67, Springer, 1991.

[10] Y. Pawitan. *In All Likelihood: Statistical Modelling and Inference Using Likelihood.* Oxford, 2001.

[11] L. J. Savage. *The Theory of Statistical Decision.* J. Am. Stat. Assoc. 46:55–67, 1951.

[12] M. J. Schervish, T. Seidenfeld, J. B. Kadane and I. Levi. *Extensions of Expected Utility Theory and Some Limitations of Pairwise Comparisons.* ISIPTA '03 (eds. Bernard, Seidenfeld and Zaffalon), 496–510, Carleton Scientific, 2003.

[13] T. A. Severini. *Likelihood Methods in Statistics.* Oxford, 2000.

[14] G. Shafer. *A Mathematical Theory of Evidence.* Princeton, 1976.

[15] G. Shafer. *Lindley's Paradox* (with discussion). J. Am. Stat. Assoc. 77:325–351, 1982.

[16] N. Shilkret. *Maxitive Measure and Integration.* Indag. Math. 33:109–116, 1971.

[17] P. Smets. *Possibilistic Inference from Statistical Data.* Proceedings of the Second World Conference on Mathematics at the Service of Man (eds. Ballester, Cardús and Trillas), 611–613, Las Palmas, 1982.

[18] A. Wald. *Contributions to the Theory of Statistical Estimation and Testing Hypotheses.* Ann. Math. Stat. 10:299–326, 1939.

[19] P. Walley. *Reconciling Frequentist Properties with the Likelihood Principle.* J. Stat. Plann. Inference 105:35–65, 2002.

[20] P. Walley and S. Moral. *Upper Probabilities Based Only on the Likelihood Function.* J. R. Stat. Soc. Ser. B Stat. Methodol. 61:831–847, 1999.