The likelihood approach to statistics as a theory of imprecise probability

Marco Cattaneo
Department of Statistics, LMU Munich
cattaneo@stat.uni-muenchen.de

September 25, 2009

likelihood function

- set $\mathcal{P}$ of probability measures on $(\Omega, \mathcal{A})$

likelihood function

- set $\mathcal{P}$ of probability measures on $(\Omega, \mathcal{A})$
- each $P \in \mathcal{P}$ is interpreted as a probabilistic model of the reality under consideration

likelihood function

- set $\mathcal{P}$ of probability measures on $(\Omega, \mathcal{A})$
- each $P \in \mathcal{P}$ is interpreted as a probabilistic model of the reality under consideration
- after having observed the event $A \in \mathcal{A}$, the **likelihood function** $lik(P) \propto P(A)$ on $\mathcal{P}$ describes the *relative* ability of the models to forecast the observed data

likelihood function

- set $\mathcal{P}$ of probability measures on $(\Omega, \mathcal{A})$
- each $P \in \mathcal{P}$ is interpreted as a probabilistic model of the reality under consideration
- after having observed the event $A \in \mathcal{A}$, the **likelihood function** $lik(P) \propto P(A)$ on $\mathcal{P}$ describes the *relative* ability of the models to forecast the observed data
- $\log \frac{lik(P_1)}{lik(P_2)}$ is the *information for discrimination* (or *weight of evidence*) in favor of P_1 against P_2

likelihood function

- set $\mathcal{P}$ of probability measures on $(\Omega, \mathcal{A})$
- each $P \in \mathcal{P}$ is interpreted as a probabilistic model of the reality under consideration
- after having observed the event $A \in \mathcal{A}$, the **likelihood function** $lik(P) \propto P(A)$ on $\mathcal{P}$ describes the *relative* ability of the models to forecast the observed data
- $\log \frac{lik(P_1)}{lik(P_2)}$ is the *information for discrimination* (or *weight of evidence*) in favor of P_1 against P_2
- in particular, a constant *lik* describes the case of **no information** for discrimination among the probabilistic models in $\mathcal{P}$

hierarchical model

- the set $\mathcal{P}$ of probability measures and the likelihood function *lik* on $\mathcal{P}$ can be interpreted as the two levels of a **hierarchical model** of the reality under consideration

hierarchical model

- the set $\mathcal{P}$ of probability measures and the likelihood function *lik* on $\mathcal{P}$ can be interpreted as the two levels of a **hierarchical model** of the reality under consideration

- when an event $A \in \mathcal{A}$ is observed, the hierarchical model can be updated as follows:

$$\begin{aligned} \mathcal{P} \quad &\rightsquigarrow \quad \mathcal{P}' = \{P(\cdot \mid A) : P \in \mathcal{P},\, P(A) > 0\} \\ lik \quad &\rightsquigarrow \quad lik'(P') \propto \sup_{P \in \mathcal{P} \,:\, P(\cdot \mid A) = P'} lik(P)\, P(A) \quad \text{on } \mathcal{P}' \end{aligned}$$

hierarchical model

- the set $\mathcal{P}$ of probability measures and the likelihood function *lik* on $\mathcal{P}$ can be interpreted as the two levels of a **hierarchical model** of the reality under consideration

- when an event $A \in \mathcal{A}$ is observed, the hierarchical model can be updated as follows:

$$\begin{aligned} \mathcal{P} \quad &\rightsquigarrow \quad \mathcal{P}' = \{P(\cdot \mid A) : P \in \mathcal{P},\ P(A) > 0\} \\ lik \quad &\rightsquigarrow \quad lik'(P') \propto \sup_{P \in \mathcal{P} \,:\, P(\cdot \mid A) = P'} lik(P)\, P(A) \quad \text{on } \mathcal{P}' \end{aligned}$$

- the **prior** likelihood function *lik* can describe the information from past observations, or subjective beliefs (interpreted as the information from *virtual* past observations)

hierarchical model

- the set $\mathcal{P}$ of probability measures and the likelihood function *lik* on $\mathcal{P}$ can be interpreted as the two levels of a **hierarchical model** of the reality under consideration

- when an event $A \in \mathcal{A}$ is observed, the hierarchical model can be updated as follows:

$$\begin{aligned} \mathcal{P} &\quad \rightsquigarrow \quad \mathcal{P}' = \{P(\cdot \,|\, A) : P \in \mathcal{P},\ P(A) > 0\} \\ \mathit{lik} &\quad \rightsquigarrow \quad \mathit{lik}'(P') \propto \sup_{P \in \mathcal{P}\,:\, P(\cdot \,|\, A) = P'} \mathit{lik}(P)\, P(A) \quad \text{on } \mathcal{P}' \end{aligned}$$

- the **prior** likelihood function *lik* can describe the information from past observations, or subjective beliefs (interpreted as the information from *virtual* past observations)

- the penalty term in penalized likelihood methods can often be interpreted as a prior *lik*

hierarchical model

- the set $\mathcal{P}$ of probability measures and the likelihood function *lik* on $\mathcal{P}$ can be interpreted as the two levels of a **hierarchical model** of the reality under consideration

- when an event $A \in \mathcal{A}$ is observed, the hierarchical model can be updated as follows:

$$\begin{aligned} \mathcal{P} \quad &\rightsquigarrow \quad \mathcal{P}' = \{P(\cdot \mid A) : P \in \mathcal{P},\, P(A) > 0\} \\ lik \quad &\rightsquigarrow \quad lik'(P') \propto \sup_{P \in \mathcal{P} \,:\, P(\cdot \mid A) = P'} lik(P)\, P(A) \quad \text{on } \mathcal{P}' \end{aligned}$$

- the **prior** likelihood function *lik* can describe the information from past observations, or subjective beliefs (interpreted as the information from *virtual* past observations)

- the penalty term in penalized likelihood methods can often be interpreted as a prior *lik*

- the choice of a prior *lik* seems better supported by intuition than the choice of a prior probability measure: in particular, a constant *lik* describes the case of no information (**complete ignorance**)

imprecise probability

- the uncertain knowledge about the value $g(P)$ of a function $g : \mathcal{P} \to \mathcal{G}$ is described by the **profile** likelihood function

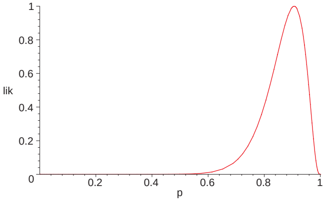
$$lik_g(\gamma) \propto \sup_{P \in \mathcal{P} \,:\, g(P) = \gamma} lik(P) \quad \text{on } \mathcal{G}$$

imprecise probability

- the uncertain knowledge about the value $g(P)$ of a function $g : \mathcal{P} \to \mathcal{G}$ is described by the **profile** likelihood function

$$lik_g(\gamma) \propto \sup_{P \in \mathcal{P} \,:\, g(P) = \gamma} lik(P) \quad \text{on } \mathcal{G}$$

- example: profile likelihood function for the probability p of observing at least 3 successes in the next 5 experiments (Bernoulli trials), after having observed 38 successes in 50 experiments

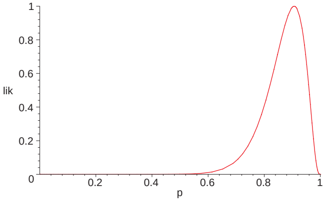


imprecise probability

- the uncertain knowledge about the value $g(P)$ of a function $g : \mathcal{P} \to \mathcal{G}$ is described by the **profile** likelihood function

$$lik_g(\gamma) \propto \sup_{P \in \mathcal{P} \,:\, g(P) = \gamma} lik(P) \quad \text{on } \mathcal{G}$$

- example: profile likelihood function for the probability p of observing at least 3 successes in the next 5 experiments (Bernoulli trials), after having observed 38 successes in 50 experiments



- *normalized* likelihood functions are a possible interpretation of membership functions of fuzzy sets: in this sense, the hierarchical model is a **fuzzy probability** measure, and the above graph shows the membership function of a fuzzy probability value

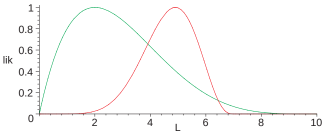
likelihood-based decisions

- a decision problem is described by a **loss function** $L : \mathcal{P} \times \mathcal{D} \to [0, \infty)$, where $L(P, d)$ is the loss incurred by making the decision d, according to the probabilistic model P

likelihood-based decisions

- a decision problem is described by a **loss function** $L : \mathcal{P} \times \mathcal{D} \to [0, \infty)$, where $L(P, d)$ is the loss incurred by making the decision d, according to the probabilistic model P

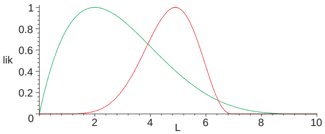
- example: profile likelihood functions for the losses $L(P, d_1)$ and $L(P, d_2)$ (i.e., membership functions for the fuzzy losses of d_1 and d_2)



likelihood-based decisions

- a decision problem is described by a **loss function** $L : \mathcal{P} \times \mathcal{D} \to [0, \infty)$, where $L(P, d)$ is the loss incurred by making the decision d, according to the probabilistic model P

- example: profile likelihood functions for the losses $L(P, d_1)$ and $L(P, d_2)$ (i.e., membership functions for the fuzzy losses of d_1 and d_2)



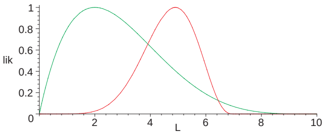
- maximum likelihood estimation leads to the MLD criterion:

$$\text{minimize} \quad L(\hat{P}_{ML}, d)$$

likelihood-based decisions

- a decision problem is described by a **loss function** $L : \mathcal{P} \times \mathcal{D} \to [0, \infty)$, where $L(P, d)$ is the loss incurred by making the decision d, according to the probabilistic model P

- example: profile likelihood functions for the losses $L(P, d_1)$ and $L(P, d_2)$ (i.e., membership functions for the fuzzy losses of d_1 and d_2)



- maximum likelihood estimation leads to the MLD criterion:

$$\text{minimize} \quad L(\hat{P}_{ML}, d)$$

- the only likelihood-based decision criterion satisfying some basic properties is the **MPL criterion** with $\alpha \in (0, \infty)$:

$$\text{minimize} \quad \sup_{P \in \mathcal{P}} lik(P)^{\alpha} \, L(P, d)$$

comparison of hierarchical and Bayesian models

- example: $\mathcal{P} = \{P_0, P_1, \ldots, P_n\}$ and $\mathcal{D} = \{d_0, d_1\}$, with
  $L(P_0, d_0) = 0$ and $L(P_i, d_0) = 1$ for all $i \in \{1, \ldots, n\}$,
  $L(P_0, d_1) = 1$ and $L(P_i, d_1) = 0$ for all $i \in \{1, \ldots, n\}$,

comparison of hierarchical and Bayesian models

- example: $\mathcal{P} = \{P_0, P_1, \ldots, P_n\}$ and $\mathcal{D} = \{d_0, d_1\}$, with
  $L(P_0, d_0) = 0$ and $L(P_i, d_0) = 1$ for all $i \in \{1, \ldots, n\}$,
  $L(P_0, d_1) = 1$ and $L(P_i, d_1) = 0$ for all $i \in \{1, \ldots, n\}$,
  - **likelihood function** *lik* on $\mathcal{P}$ with $lik(P_0) = c\, lik(P_i)$ for a $c > 1$ and all $i \in \{1, \ldots, n\}$:
    likelihood-based decision criterion $\Rightarrow$ d_0 optimal

comparison of hierarchical and Bayesian models

- example: $\mathcal{P} = \{P_0, P_1, \ldots, P_n\}$ and $\mathcal{D} = \{d_0, d_1\}$, with
  $L(P_0, d_0) = 0$ and $L(P_i, d_0) = 1$ for all $i \in \{1, \ldots, n\}$,
  $L(P_0, d_1) = 1$ and $L(P_i, d_1) = 0$ for all $i \in \{1, \ldots, n\}$,
  - **likelihood function** *lik* on $\mathcal{P}$ with $lik(P_0) = c\, lik(P_i)$ for a $c > 1$ and all $i \in \{1, \ldots, n\}$:
    likelihood-based decision criterion $\Rightarrow$ d_0 optimal
  - **probability measure** π on $\mathcal{P}$ with $\pi\{P_0\} = c\, \pi\{P_i\}$ for a $c > 1$ and all $i \in \{1, \ldots, n\}$:
    Bayesian decision criterion $\Rightarrow$ d_1 optimal when n is large enough
    (*many bad probabilistic models make a good one*)

comparison of hierarchical and Bayesian models

- example: $\mathcal{P} = \{P_0, P_1, \dots, P_n\}$ and $\mathcal{D} = \{d_0, d_1\}$, with
  $L(P_0, d_0) = 0$ and $L(P_i, d_0) = 1$ for all $i \in \{1, \dots, n\}$,
  $L(P_0, d_1) = 1$ and $L(P_i, d_1) = 0$ for all $i \in \{1, \dots, n\}$,
  - **likelihood function** *lik* on $\mathcal{P}$ with $lik(P_0) = c\, lik(P_i)$ for a $c > 1$ and all $i \in \{1, \dots, n\}$:
    likelihood-based decision criterion $\Rightarrow$ d_0 optimal
  - **probability measure** π on $\mathcal{P}$ with $\pi\{P_0\} = c\, \pi\{P_i\}$ for a $c > 1$ and all $i \in \{1, \dots, n\}$:
    Bayesian decision criterion $\Rightarrow$ d_1 optimal when n is large enough
    (*many bad probabilistic models make a good one*)

- in the Bayesian approach the probabilistic models are handled as possible "states of the world" (in particular, they are considered *mutually exclusive*)

comparison of hierarchical and Bayesian models

- example: $\mathcal{P} = \{P_0, P_1, \ldots, P_n\}$ and $\mathcal{D} = \{d_0, d_1\}$, with
  $L(P_0, d_0) = 0$ and $L(P_i, d_0) = 1$ for all $i \in \{1, \ldots, n\}$,
  $L(P_0, d_1) = 1$ and $L(P_i, d_1) = 0$ for all $i \in \{1, \ldots, n\}$,
  - **likelihood function** *lik* on $\mathcal{P}$ with $lik(P_0) = c\, lik(P_i)$ for a $c > 1$ and all $i \in \{1, \ldots, n\}$:
    likelihood-based decision criterion $\Rightarrow$ d_0 optimal
  - **probability measure** π on $\mathcal{P}$ with $\pi\{P_0\} = c\, \pi\{P_i\}$ for a $c > 1$ and all $i \in \{1, \ldots, n\}$:
    Bayesian decision criterion $\Rightarrow$ d_1 optimal when n is large enough
    (*many bad probabilistic models make a good one*)

- in the Bayesian approach the probabilistic models are handled as possible "states of the world" (in particular, they are considered *mutually exclusive*)

- basic advantage of the hierarchical model over

comparison of hierarchical and Bayesian models

- example: $\mathcal{P} = \{P_0, P_1, \ldots, P_n\}$ and $\mathcal{D} = \{d_0, d_1\}$, with
  $L(P_0, d_0) = 0$ and $L(P_i, d_0) = 1$ for all $i \in \{1, \ldots, n\}$,
  $L(P_0, d_1) = 1$ and $L(P_i, d_1) = 0$ for all $i \in \{1, \ldots, n\}$,
  - **likelihood function** *lik* on $\mathcal{P}$ with $lik(P_0) = c\, lik(P_i)$ for a $c > 1$ and all $i \in \{1, \ldots, n\}$:
    likelihood-based decision criterion $\Rightarrow$ d_0 optimal
  - **probability measure** π on $\mathcal{P}$ with $\pi\{P_0\} = c\, \pi\{P_i\}$ for a $c > 1$ and all $i \in \{1, \ldots, n\}$:
    Bayesian decision criterion $\Rightarrow$ d_1 optimal when n is large enough
    (*many bad probabilistic models make a good one*)

- in the Bayesian approach the probabilistic models are handled as possible "states of the world" (in particular, they are considered *mutually exclusive*)

- basic advantage of the hierarchical model over
  - the precise Bayesian model: the ability to describe the state of **complete ignorance**

comparison of hierarchical and Bayesian models

- example: $\mathcal{P} = \{P_0, P_1, \ldots, P_n\}$ and $\mathcal{D} = \{d_0, d_1\}$, with
  $L(P_0, d_0) = 0$ and $L(P_i, d_0) = 1$ for all $i \in \{1, \ldots, n\}$,
  $L(P_0, d_1) = 1$ and $L(P_i, d_1) = 0$ for all $i \in \{1, \ldots, n\}$,
  - **likelihood function** *lik* on $\mathcal{P}$ with $lik(P_0) = c\, lik(P_i)$ for a $c > 1$ and all $i \in \{1, \ldots, n\}$:
    likelihood-based decision criterion $\Rightarrow$ d_0 optimal
  - **probability measure** π on $\mathcal{P}$ with $\pi\{P_0\} = c\,\pi\{P_i\}$ for a $c > 1$ and all $i \in \{1, \ldots, n\}$:
    Bayesian decision criterion $\Rightarrow$ d_1 optimal when n is large enough
    (*many bad probabilistic models make a good one*)

- in the Bayesian approach the probabilistic models are handled as possible "states of the world" (in particular, they are considered *mutually exclusive*)

- basic advantage of the hierarchical model over
  - the precise Bayesian model: the ability to describe the state of **complete ignorance**
  - the imprecise Bayesian model: the ability to **get out** of the state of complete ignorance