The likelihood approach to statistical decision problems

Marco Cattaneo
Department of Statistics, LMU Munich

27 January 2014

notation

- in statistics, L usually denotes:
  - likelihood function (here λ)
  - loss function (here W)
- **statistical model**: $(\Omega, \mathcal{F}, P_\theta)$ with $\theta \in \Theta$ (where Θ is a nonempty set) and random variables $X : \Omega \to \mathcal{X}$ and $X_n : \Omega \to \mathcal{X}_n$

loss function

- a statistical **decision problem** is described by a loss function

  $$W : \Theta \times \mathcal{D} \to [0, +\infty[,$$

  where $\mathcal{D}$ is a nonempty set

- intended as unification (and generalization) of statistical inference, in particular of:
  - point estimation (e.g., with $\mathcal{D} = \Theta$)
  - hypothesis testing (e.g., with $\mathcal{D} = \{H_0, H_1\}$)

- most successful general methods:
  - point estimation: maximum likelihood estimators
  - hypothesis testing: likelihood ratio tests

- these methods do not fit well in the setting of statistical decision theory: here they are unified (and generalized) in likelihood decision theory

likelihood function

- $\lambda_x : \Theta \to [0, 1]$ is the (relative) likelihood function given $X = x$, when
$$\sup_{\theta \in \Theta} \lambda_x(\theta) = 1 \quad \text{and} \quad \lambda_x(\theta) \propto P_\theta(X = x)$$
(with $\lambda_x(\theta) \propto f_\theta(x)$ as approximation for continuous X)

- λ_x describes the relative plausibility of the possible values of θ in the light of the observation $X = x$, and can thus be used as a basis for post-data decision making

- prior information can be described by a prior likelihood function: if X_1 and X_2 are independent, then $\lambda_{(x_1,x_2)} \propto \lambda_{x_1} \lambda_{x_2}$; that is, when $X_2 = x_2$ is observed, the prior λ_{x_1} is updated to the posterior $\lambda_{(x_1,x_2)}$

- strong similarity with the Bayesian approach (both satisfy the likelihood principle): a fundamental advantage of the likelihood approach is the possibility of not using prior information (since $\lambda_{x_1} \equiv 1$ describes complete ignorance)

likelihood decision criteria

- likelihood decision criterion: minimize $V(W(\cdot, d), \lambda_x)$, where the functional V must satisfy the following three properties, for all functions $w, w' : \Theta \to [0, +\infty[$ and all likelihood functions $\lambda, \lambda_n : \Theta \to [0, 1]$
  - monotonicity: $w \leq w'$ (pointwise) $\Rightarrow V(w, \lambda) \leq V(w', \lambda)$
    (implied by meaning of W)
  - parametrization invariance: $b : \Theta \to \Theta$ bijection $\Rightarrow V(w \circ b, \lambda \circ b) = V(w, \lambda)$
    (excludes Bayesian criteria $V(w, \lambda) = \frac{\int w \, \lambda \, d\mu}{\int \lambda \, d\mu}$ for infinite Θ)
  - consistency: $\mathcal{H} \subseteq \Theta$ with $\lim_{n \to \infty} \sup_{\theta \in \Theta \setminus \mathcal{H}} \lambda_n(\theta) = 0 \Rightarrow$
    $\lim_{n \to \infty} V(c \, I_{\mathcal{H}} + c' \, I_{\Theta \setminus \mathcal{H}}, \lambda_n) = c$ for all constants $c, c' \in [0, +\infty[$
    (excludes minimax criterion $V(w, \lambda) = \sup_{\theta \in \Theta} w(\theta)$,
    implies calibration: $V(c, \lambda) = c$)

- **likelihood decision function**: $\delta : \mathcal{X} \to \mathcal{D}$ such that $\delta(x)$ minimizes $V(W(\cdot, d), \lambda_x)$

properties

- likelihood decision criteria have the advantages of **post-data** methods:
  - independence from choice of possible alternative observations
  - direct interpretation
  - simpler problems
- likelihood decision criteria have also important **pre-data** properties:
  - equivariance: for invariant decision problems, the likelihood decision functions are equivariant
  - (strong) consistency: under some regularity conditions, the likelihood decision functions $\delta_n : \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \to \mathcal{D}$ satisfy

$$\lim_{n \to \infty} W(\theta, \delta_n(X_1, \ldots, X_n)) = \inf_{d \in \mathcal{D}} W(\theta, d) \quad P_\theta\text{-a.s.}$$

MPL criterion

- MPL criterion: minimize $\sup_{\theta \in \Theta} W(\theta, d)\, \lambda_x(\theta)$, corresponds to
  $$V(w, \lambda) = \sup_{\theta \in \Theta} w(\theta)\, \lambda(\theta)$$
  (nonadditive integral of w with respect to $\mathcal{H} \mapsto \sup_{\theta \in \mathcal{H}} \lambda(\theta)$)

- point estimation:
  - $\mathcal{D} = \Theta$ finite
  - $W(\theta, \hat{\theta}) = I_{\theta \neq \hat{\theta}}$ simple loss function
  - the **maximum likelihood estimator** (when well-defined) is the likelihood decision function resulting from the MPL criterion

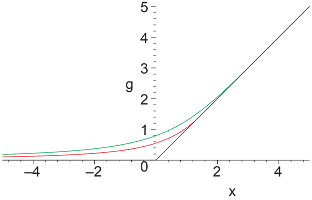
- hypothesis testing:
  - $\mathcal{D} = \{H_0, H_1\}$ with $H_0 : \theta \in \mathcal{H}$ and $H_1 : \theta \in \Theta \setminus \mathcal{H}$
  - $W(\theta, H_1) = c\, I_{\theta \in \mathcal{H}}$ and $W(\theta, H_0) = c'\, I_{\theta \in \Theta \setminus \mathcal{H}}$ with $c \geq c'$
  - the **likelihood ratio test** with critical value c'/c is the likelihood decision function resulting from the MPL criterion

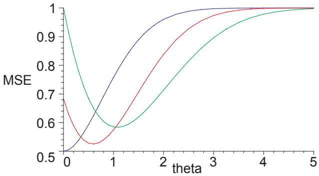
a simple example

- $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\theta, \sigma^2)$ with $\Theta =]0, +\infty[$ (that is, θ positive and σ known)

- estimation of θ with squared error:
  - $\mathcal{D} = \Theta$ with $W(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$
  - no unbiased estimator, maximum likelihood estimator not well-defined, no standard (proper) Bayesian prior

- likelihood decision function resulting from the **MPL criterion**:
  - scale invariance and sufficiency: $\hat{\theta}(x_1, \ldots, x_n) = g\left(\frac{\bar{x}}{\sigma/\sqrt{n}}\right) {}^{\sigma}/_{\sqrt{n}}$
  - consistency and asymptotic efficiency: $\hat{\theta}(x_1, \ldots, x_n) = \bar{x}$ when $\bar{x} \geq {}^{\sqrt{2}\,\sigma}/_{\sqrt{n}}$





conclusion

- this work:
  - fills a gap in the likelihood approach to statistics
  - introduces an alternative to classical and Bayesian decision making
  - offers a new perspective on the likelihood methods
- **likelihood decision making**:
  - is post-data and equivariant
  - is consistent and asymptotically efficient
  - does not need prior information

references

- Lehmann (1959). **Testing Statistical Hypotheses**. Wiley.
- Diehl and Sprott (1965). **Die Likelihoodfunktion und ihre Verwendung beim statistischen Schluß**. *Statistische Hefte* 6, 112–134.
- Giang and Shenoy (2005). **Decision making on the sole basis of statistical likelihood**. *Artificial Intelligence* 165, 137–163.
- Cattaneo (2013). **Likelihood decision functions**. *Electronic Journal of Statistics* 7, 2924–2946.
- Cattaneo and Wiencierz (2012). **Likelihood-based Imprecise Regression**. *International Journal of Approximate Reasoning* 53, 1137–1154.