Lecture 3: Decision theory

Sébastien Gadat

TSE

November 7, 2022

Decision theory

History : Wald (1939), Lehman (1950), Savage (1954)

Let $(\Omega, \mathcal{P})$ be a parametric model with

$$\mathcal{P} = \{\mathbb{P}_\theta;\, \theta \in \Theta\}.$$

Belief : the DGP ("data generating process") is the mechanism which generates the data from the law $\mathbb{P}_{\theta_0}$ where the value θ_0 is called the "truth", or "state of nature"

Objective : guess the truth about the DGP (*i.e.* estimate and make inference on θ_0) using the available observed data.

Decision rule

Depending upon the objectives of the statistician, one defines a **set of possible decisions** $\mathcal{D}$**.** What is a decision ?

A decision corresponds to some statement relative to θ.

The statistician must take one decision $d \in \mathcal{D}$ based on one observation $x \in \Omega$.

Definition of a decision rule : It is a map r which assigns exactly one decision $d \in \mathcal{D}$ to each observation $x \in \Omega$:

$$\begin{aligned} r \quad : \quad & \Omega \to \mathcal{D} \\ & x \mapsto r(x) = d. \end{aligned}$$

A decision rule defines a strategy for the statistician.
We will denote by $\mathcal{R}$ the set of decision rules r.

Example 1

We experiment a medical treatment on a population of n sick patients.

Let θ be the true proportion of patients cured by this treatment.

We cannot do the experiment on the whole population (too expensive), and hence we cannot observe the true θ, but only an approximation $\hat{\theta}$ of θ based on the available sample of sick patients.

We would like to solve questions relative to θ based on the proportion $\hat{\theta}$ of patients cured by the treatment in the sample.

Which model to use ? If we associate to each patient i the r.v. $X_i = 1$ if cured, and $X_i = 0$ otherwise, then

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} X_i \quad \text{and} \quad n\hat{\theta} \rightsquigarrow \mathcal{B}(n, \theta) .$$

Example 1

- Objective 1 : estimate the proportion θ of cured patients, we have an estimation problem and a decision is an estimate $\hat{\theta}$ in $\mathcal{D} = [0,1] \equiv \Theta$.
- Objective 2 : decide **whether or not** the proportion θ is larger than a given threshold θ^* (could be the proportion of cured patients with a former treatment), we have a test problem and there are two possible decisions (yes or no, *i.e.*, $\mathcal{D} = \{0,1\}$).
- Objective 3 : give a bracket $[\theta_{low}, \theta_{high}]$ such that the true θ belongs to this bracket with a high probability (confidence level), we have a CI problem and a decision $d = r(x)$ is an interval with bounds in $[0,1]$.

Types of decision problems

1. Point estimation problem : $\mathcal{D} = \Theta$ and $r(x) = \hat{\theta}$.
2. Test problem : $\mathcal{D} = \{0, 1\}$, $r(x) = 0$ if $\theta \in H_0$ and $r(x) = 1$ if $\theta \in H_1$: a decision rule is therefore an indicator function
3. Confidence interval problem : $\mathcal{D}$ is the set of intervals with bounds in Θ and $r(x) = [\hat{LB}, \hat{UB}]$.

If r is a decision rule and x is a realization of X, then $r(x)$ is a realization of the random variable $r(X)$.

For a same problem, one can define several decision rules r (*e.g.* one may construct different estimates $\hat{\theta}$ of θ from the same data x).

Loss function

Loss function ? : it is a map

$$L : \Theta \times \mathcal{D} \mapsto \mathbb{R}^+$$

which assigns a non negative real number to each pair (θ, d) where $\theta \in \Theta$ and $d \in \mathcal{D}$ is a decision.

- $L(\theta, d)$ measures the size of a possibly wrong decision d (when the true probability law corresponds to the parameter θ).
- the choice of L depends on the decision problem and is left to the statistician or the decision maker, it is not dictated by the experiment.

Loss function : examples

For estimation problems :

- the most frequent loss function is the quadratic loss

$$L(\theta, d) = (\theta - d)^2 \quad \text{if} \quad \theta,\ d \in \mathbb{R};$$

- an alternative is the absolute loss

$$L(\theta, d) = \mid \theta - d \mid .$$

Loss function : examples

For a test problem, assume we test H_0 against H_1. There are two possible decisions and two possibilities for θ (it satisfies H_0 or not), therefore there are only four losses to define.

truth $\mapsto$ decision $\downarrow$	H_0	H_1
H_0	$L = 0$	$L = L_2$
H_1	$L = L_1$	$L = 0$

L_1 is the first type error, i.e. reject H_0 whereas it is true
L_2 is the second type error, i.e. do not reject H_0 whereas it is wrong

Risk function

If d is a decision rule, then $L(\theta, d(X))$ is a random variable (where X follows P_θ).

In order to use this r.v. for choosing among decision rules, one needs to summarize this random loss by taking its expectation, which yields the **risk function** (risk of taking a possibly wrong or inaccurate decision rule) :

$$R(\theta, d) = \mathbb{E}_\theta[L(\theta, d(X))].$$

Meaning of $\mathbb{E}_\theta$: expected value when X follows the distribution $\mathbb{P}_\theta$ corresponding to the value θ of the model parameter.

We will write also $\mathbb{E}_X$ later with the same meaning when there is no possible confusion about the value of θ.

Decomposition of quadratic risk

For quadratic loss :

$$L(\theta, d) = (\theta - d)^2 \quad \text{if} \quad \theta,\, d \in \mathbb{R}$$

there is a classical decomposition of quadratic risk for an estimator $d(X) = \hat{\theta}$ of a real parameter θ.

Theorem (Bias Variance decomposition)

$$\begin{aligned} R(\theta, d) &= \mathbb{E}_\theta[(\theta - \hat{\theta})^2] \\ &= [\mathbb{E}_\theta(\hat{\theta}) - \theta]^2 + var_\theta \\ &= squared\ bias + variance \end{aligned}$$

Examples of Risk computation

- Let $(X_1, \ldots, X_n)$ be a random sample from a Poisson distribution with mean λ.

  Using quadratic loss, evaluate the risk of the estimator $\hat{\lambda} = \bar{X}$.

- Let $(X_1, \ldots, X_n)$ be a random sample from a Bernoulli distribution with parameter p.

  Using quadratic loss, evaluate the risk of the estimator $\hat{p} = \bar{X}$.

Evaluating risk

- Exact finite distance computation (school exercise as seen in the previous slide, but seldom possible in practice).
- Asymptotic evaluation (when the sample size is large enough).
- Bootstrap (see the prinicple in the next slide).

Bootstrap principle

Purpose of bootstrap : estimate bias, variance and mean squared error of estimators $\hat{\theta} = T(X_1, \ldots, X_n)$ when no closed formula is possible.

The general idea of the method is to use the initial sample $X_1, \ldots, X_n$ to draw B samples, called bootstrap samples and denoted $X_1^{*(b)}, \ldots, X_n^{*(b)}$, and to use these generated samples to evaluate the risk.

- In the real world, a **single** sample implies a **single** estimate $\hat{\theta} = T(X_1, \ldots, X_n)$ of the unknown parameter θ, hence it is impossible to estimate bias and variance.
- In the bootstrap world (conditionally on the initial sample), what plays the role of the unknown θ is $\hat{\theta}$. We can compute B estimates of θ by $\hat{\theta}^{*(b)} = T(X_1^{*(b)}, \ldots, X_n^{*(b)})$, and their sample average is

$$\bar{\hat{\theta}}^{*} = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^{*(b)} \quad \text{which estimates} \quad \mathbb{E}_\theta(\hat{\theta}).$$

Two kinds of bootstrap

The way these bootstrap samples are generated differs according to the type of bootstrap :

- **nonparametric** : draw B independent bootstrap samples of size n, denoted by $X_1^{*(b)}, \ldots, X_n^{*(b)}$, from the empirical cumulative distribution function of the initial sample.

- **parametric** :
  1. choose a parametric family of distributions ;
  2. fit this model to the initial sample to get parameters estimates ;
  3. draw B independent bootstrap samples of size n, denoted by $X_1^{*(b)}, \ldots, X_n^{*(b)}$, from the fitted distribution obtained in step 2.

Bootstrap : risk evaluation

1. to estimate the bias $(\mathbb{E}_\theta(\hat{\theta}) - \theta)$ of $\hat{\theta} = T(X_1, \ldots, X_n)$, first compute $\bar{\hat{\theta}}^* = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}^{*(b)}$ and then

$$\widehat{\text{Bias}} = \bar{\hat{\theta}}^* - \hat{\theta}.$$

2. to estimate the variance $\mathbb{E}_\theta\left[\left(\hat{\theta} - \mathbb{E}_\theta(\hat{\theta})\right)^2\right]$ of $\hat{\theta}$, use

$$\widehat{\text{Var}} = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}^{*(b)} - \bar{\hat{\theta}}^*\right)^2.$$

3. to estimate $\text{MSE} = \mathbb{E}_\theta\left[(\hat{\theta} - \theta)^2\right] = \text{squared bias} + \text{variance}$, use

$$\widehat{MSE} = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}^{*(b)} - \hat{\theta}\right)^2 = (\widehat{\text{Bias}})^2 + \widehat{\text{Var}}.$$

Exercise : prove that the two formulas for $\widehat{MSE}$ are equivalent.

Comparing decision rules using the risk function

- The statistician will try to minimize the risk, but note that $R(\theta, d)$ is a function of the unknown parameter θ.
- Two functions $\theta \mapsto R(\theta, d_1)$ and $\theta \mapsto R(\theta, d_2)$, may not be comparable...

Preference relations :

A decision d_1 **is preferred to** d_2 if its risk is entirely below that of d_2 :

$$\forall \theta \qquad R(\theta, d_1) \leq R(\theta, d_2)$$

A decision d_1 **is strictly preferred to** d_2 if

$$\forall \theta \qquad R(\theta, d_1) \leq R(\theta, d_2) \qquad \text{and} \qquad \exists \theta^* \quad R(\theta^*, d_1) < R(\theta^*, d_2).$$

Comparing decision rules using the risk function

The preference relation is a partial order : there exist rules which are not comparable.

Example : $\Omega = \{0, 1\}$, and $\mathcal{P}$ is the set of Bernoulli distributions with parameter $\theta \in [0, 1]$.

Assume $\mathcal{D} = \{0, 1\}$ and $L(\theta, d) = (1 - \theta) \cdot 1_{(d=1)} + \theta \cdot 1_{(d=0)}$
Given an observation X, four possible decision rules :

	d_1	d_2	d_3	d_4
If $X = 0$	0	1	0	1
If $X = 1$	0	0	1	1

We can compute the risks !

$$R(\theta, d_1) = \theta, \quad R(\theta, d_4) = 1 - \theta,$$

$$R(\theta, d_2) = (1 - \theta)^2 + \theta^2, \quad R(\theta, d_3) = 2\theta(1 - \theta).$$

$\hookrightarrow$ d_3 is preferred to d_2, $\hookrightarrow$ d_1 and d_4 are not comparable.

Optimal decision rule

It is tempting to define an ideal decision rule by the following :

Definition : An **optimal** decision rule is a rule which is preferred to any other rule.

However in general, there exists no such rule !!
The set of rules is too large implying too many constraints.

A more realistic requirement that an acceptable decision rule must fulfill is the following :

Definition : A decision rule d is **admissible** if there exists no decision which is strictly preferred to d.

If d is admissible, it does not mean that d is preferred to any other rule.

Principles to decide between decision rules

- The preference relation is partial : there exist rules that are not comparable
- The optimality is useless : typically there exists no decision which is preferred to any other rule

Consequence : the statistician must resort to principles such as :

- minimax principle
- bayesian principle
- optimality in a subclass
- maximum likelihood
- asymptotics

Minimax principle

Goal : Summarize the risk in a unique number.
Tool : R_M : the minimax risk of a decision rule d is the highest risk $R(\theta, d)$ when θ varies in the space of parameters Θ.

Definition d^* is a minimax rule in a given class of rules $\mathcal{D}_1$ if its minimax risk is minimum in this class :

$$R_M(d^*) = \inf_{d \in \mathcal{D}_1} \sup_{\theta \in \Theta} R(\theta, d)$$

Remark :

- $\mathcal{D}_1$ may be a strict subset of $\mathcal{D}$.
- d^* is trying to do as well as possible in the worst case, which is somewhat sad ;)

Minimax principle : example

$\Omega = \{0, 1\}$ and $\mathcal{P}$ is the set of bernoulli distributions with parameter $\theta \in [0, 1]$. Let $X_1, \ldots, X_n$ be a sample from this model and $\hat{\theta}_1 = \bar{X}$ be the sample mean.

Prove that with the squared error loss :

$$R(\theta, \bar{X}) = \frac{\theta(1-\theta)}{n},$$

$$R_M(\bar{X}) = \frac{1}{4n}.$$

Let

$$\hat{\theta}_2 = \frac{n\bar{X} + \frac{\sqrt{n}}{2}}{n + \sqrt{n}}.$$

Prove that

$$R_M(\hat{\theta}_2) = \frac{1}{4(\sqrt{n}+1)^2}$$

and therefore that $\bar{X}$ is not minimax.

Review on conditional expectation

Given a pair of r.v. (T, X), the conditional expectation denoted by :

$$\mathbb{E}(X|T)$$

is a measurable function of T and satisfies :

1. minimizes $\mathbb{E}[X - f(T)]^2$ when f varies in the measurable functions.
2. $\mathbb{E}(X|T = t)$: mean of the conditional distribution of X given $T = t$.

Interpretation : $\mathbb{E}(X|T)$ is the closest r.v. to X in terms of the mean squared deviation $\mathbb{E}[X - f(T)]^2$.

Example : if (T, X) is continuous, then :

$$\mathbb{E}(X|T = t) = \int_{-\infty}^{+\infty} x\, f_{X|T}(x \mid t)dx = \int_{-\infty}^{+\infty} x\, \frac{f_{(T,X)}(t, x)}{f_T(t)}dx$$

if (T, X) is discrete, then :

$$\mathbb{E}(X|T = t) = \sum_{x \in \mathcal{X}} x\, \mathbb{P}(X = x \mid T = t) = \sum_{x \in \mathcal{X}} x\, \frac{\mathbb{P}(X = x, T = t)}{\mathbb{P}(T = t)}.$$

Law of total expectation

For a measurable function h,

$$\mathbb{E}_{X,T}[h(X,T)] = \mathbb{E}_T\left(\mathbb{E}_{X|T}\left[h(X,T) \mid T\right]\right).$$

In this equality, on the l-h.s the **expectation** is with respect to the joint distribution of X and T.

The first **expectation** on the r-h.s is based on the distribution of T, and the second expectation is based on the conditional distribution of X given T.

When $h(X,T) = X$, the law of total expectation reads as :

$$\mathbb{E}_X(X) = \mathbb{E}_T\left(\mathbb{E}_{X|T}\left[X \mid T\right]\right).$$

Bayesian principle

Bayesian statistics try to introduce **expert knowledge** in the decision.

To represent **expert knowledge**, the parameter is modeled as random with a distribution called the **prior** distribution.

If the parameter of interest is λ (denoted θ before), we will denote by Λ the corresponding r.v. for which λ is a realization (it is hypothesized that the true parameter λ is drawn from the expert distribution of Λ).

In this context, $\mathbb{P}_\lambda$ is not any more the distribution $\mathbb{P}_X$ of the r.v. of interest X, but is the conditional distribution $\mathbb{P}_{X|\Lambda=\lambda}$ of X **given the value of the parameter** $\Lambda = \lambda$.

Similarly, the risk $R(\lambda, d)$ is interpreted as the conditional risk of d given that $\Lambda = \lambda$:

$$R(\lambda, d) = \mathbb{E}_{X|\Lambda}[L(\lambda, d(X))|\Lambda = \lambda].$$

Bayesian principle

In order to summarize the risk in a unique number, we define the bayesian risk $R_B(d)$ of a decision rule d for the prior law of Λ as :

$$R_B(d) = \mathbb{E}_\Lambda(R(\Lambda, d)) = \mathbb{E}_\Lambda\left(\mathbb{E}_{X|\Lambda}[L(\Lambda, d(X))|\Lambda]\right).$$

The law of total expectation leads to :

$$R_B(d) = \mathbb{E}_{X,\Lambda}[L(\Lambda, d(X))],$$

where the expectation is with respect to the joint law of X and Λ.

Definition d^* is a bayesian rule for the prior law of Λ if its bayesian risk is minimum.

Bayesian principle

A posterior distribution : it is the law of Λ given the observation $X = x$.

It reflects an update of the prior Λ through the information $X = x$.

Mathematically, one obtains the posterior by the Bayes formula.

In the continuous case, the posterior density is given by :

$$f(\lambda \mid x) = \frac{f_{\Lambda,X}(\lambda, x)}{f_X(x)},$$

where the marginal itself $f_X(x)$ is obtained as

$$f_X(x) = \int f_{\Lambda,X}(\lambda, x) d\lambda.$$

Computation of bayesian rule for quadratic loss

Theorem. When the set of parameters is an interval of $\mathbb{R}$ and the loss is quadratic, the bayesian estimator of λ is given by the posterior mean

$$r^*(x) = \mathbb{E}(\Lambda \mid X = x).$$

(see slide 28 for the definition of a bayesian rule r^*)

To apply the theorem one must follow these steps :

- the joint law of X and Λ (multiply the conditional distribution of X given Λ by the prior of Λ)
- the marginal law of X (integrate the joint wrt λ)
- the conditional law of Λ given X (posterior law)
- the mean of this conditional law (posterior mean)

Example of Bayesian estimator for quadratic loss

$X \sim \mathcal{B}(n, \lambda)$, $\lambda \in [0, 1]$. A priori law of λ : Beta distribution(a,b), its pdf

$$g(\lambda) = \lambda^{a-1}(1-\lambda)^{b-1}/B(a,b)$$

where $B(a,b) = \int_0^1 u^{a-1}(1-u)^{b-1}du$. Mean of this Beta : $\frac{a}{a+b}$.

Show that the bayesian estimator of λ under squared error loss is given by

$$\hat{\lambda} = \frac{n}{n+a+b}\frac{X}{n} + \frac{a+b}{n+a+b}\frac{a}{a+b}$$

Weighted average of the empirical mean $\frac{X}{n}$ and the a priori mean $\frac{a}{a+b}$: this shows how the bayesian principle combines prior information and data.
It is not always the case !

Example of Bayesian estimator for non quadratic loss

$X \sim \mathcal{B}(n, \lambda)$, $\lambda \in [0, 1]$. A priori law of λ : uniform on $[0, 1]$.

- Find the bayesian estimator of λ under the square error loss
- Same question with the following loss function :

$$L(\lambda, d) = \frac{(\lambda - d)^2}{\lambda(1 - \lambda)}.$$

Discussion of Bayesian rules

1. Pro : construction of estimators with good properties (most bayesian rules are admissible)

2. Cons1 : the choice of the a priori knowledge is often criticized (more driven by mathematical convenience than by expert knowledge)
3. Cons2 : the posterior distribution may not be explicit, and we shall need some simulation algorithms to compute it. See Lecture : " Stochastic methods for Optimization and Sampling."

Optimality in a subclass

To relax the optimality condition, we can restrict the set of rules to a subclass. It is usual to use the class of unbiased rules.

Preliminary notation : because the true distribution $\mathbb{P}_\theta$ is unknown, the expectation symbol will be indexed by the true value of the parameter : $\mathbb{E}_\theta$

Definition : A decision d is unbiased if for any two values of the parameter θ_1 and θ_2 we have

$$\mathbb{E}_{\theta_1}[L(\theta_1, d(X))] \leq \mathbb{E}_{\theta_1}[L(\theta_2, d(X))]$$

It means that for any value of the true parameter θ_1, the decision $d(X)$ is closer on average to the true decision θ_1 than to any other erroneous decision corresponding to θ_2. Note that this applies to the point estimation problem as well as to the test problem.

Optimality in a subclass

Theorem In the case of the point estimation problem with squared error loss, an estimator $\hat{\theta}$ such that $\mathbb{E}_\theta(\hat{\theta})$ belongs to the set of parameters Θ is unbiased if and only if for all $\theta \in \Theta$:

$$\mathbb{E}_\theta(\hat{\theta}) = \theta.$$

This leads us to the definition of optimality in the subclass of unbiased decision rules :

Definition An estimator is optimal in the class of unbiased estimators if it is preferred to any other estimator (in the sense described in slide 19).

Theorem In the case of the point estimation problem with squared error loss, an estimator $\hat{\theta}$ is optimal in the class of unbiased estimators if and only if it has Minimal Variance among Unbiased Estimators for any value of the true parameter.
This theorem explains the acronym MVUE

Optimality in a subclass : Gauss-Markov theorem

Linear model

$$Y = X\beta + \epsilon$$

Conditions : ϵ are centered and independent r.v. with bounded variance.

1. Prove that the OLS estimator of β is given by

$$\hat{\beta} = (X'X)^{-1}X'Y.$$

   It is a linear function of Y.
2. Show that the OLS estimator $\hat{\beta}$ of β is the best estimator in the subclass of linear unbiased estimators under squared error loss.

Optimality in a subclass : limitation

Even if an estimator is MVUE it may not be admissible !

Example : In the model with a random sample from $\mathcal{P}$ the set of gaussian random variables with mean zero and variance σ^2 ($\sigma > 0$), let $\theta = \sigma^2$ and

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} X_i^2.$$

1. Prove that $\hat{\theta}$ is unbiased.
2. Show that there exists a positive real α such that $\hat{\theta}_\alpha = \alpha\hat{\theta}$ has a smaller MSE than that of $\hat{\theta}$.