Econ 2140, spring 2018, Part IIa
Statistical Decision Theory

Maximilian Kasy

Department of Economics, Harvard University

Examples of decision problems

- Decide whether or not the hypothesis of no racial discrimination in job interviews is true
- Provide a forecast of the unemployment rate next month
- Provide an estimate of the returns to schooling
- Pick a portfolio of assets to invest in
- Decide whether to reduce class sizes for poor students
- Recommend a level for the top income tax rate

Takeaways for this part of class

1. A general framework to think about what makes a “good” estimator, test, etc.
2. How the foundations of statistics relate to those of microeconomic theory.
3. In what sense the set of Bayesian estimators contains most “reasonable” estimators.

Textbooks

- **Robert, C. (2007).** *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Verlag, chapter 2.
- **Casella, G. and Berger, R. (2001).** *Statistical inference*. Duxbury Press, chapter 7.3.4.

Roadmap

- IIa
  - Basic definitions
  - Optimality criteria
- IIb
  - Relationships between optimality criteria
  - Analogies to microeconomics
  - Two justifications of the Bayesian approach
  - Testing and the Neyman Pearson lemma
- IIc
  - Value added estimation
  - Ridge regression and Lasso
  - Experimental design

Part IIIa

Components of a general statistical decision problem

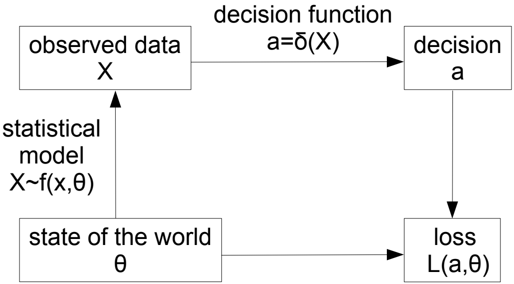
- Observed data X
- A statistical decision a
- A state of the world θ
- A loss function $L(a, \theta)$ (the negative of utility)
- A statistical model $f(X|\theta)$
- A decision function $a = \delta(X)$

How they relate

- underlying state of the world θ
  $\Rightarrow$ distribution of the observation X.
- decision maker: observes $X \Rightarrow$ picks a decision a
- her goal: pick a decision that minimizes loss $L(a, \theta)$
  (θ unknown state of the world)
- X is useful $\Leftrightarrow$ reveals some information about θ
  $\Leftrightarrow f(X|\theta)$ does depend on θ.
- problem of statistical decision theory:
  find decision functions δ which "make loss small."

Graphical illustration

Figure: A general decision problem



Examples

- investing in a portfolio of assets:
  - X: past asset prices
  - a: amount of each asset to hold
  - θ: joint distribution of past and future asset prices
  - L: minus expected utility of future income
- decide whether or not to reduce class size:
  - X: data from project STAR experiment
  - a: class size
  - θ: distribution of student outcomes for different class sizes
  - L: average of suitably scaled student outcomes, net of cost

Practice problem

For each of the examples on slide 2, what are

- the data X,
- the possible actions a,
- the relevant states of the world θ, and
- reasonable choices of loss function L?

Loss functions in estimation

- goal: find an a
- which is close to some function μ of θ.
- for instance: $\mu(\theta) = E[X]$
- loss is larger if the difference between our estimate and the true value is larger

Some possible loss functions:

1. **squared error** loss,

$$L(a, \theta) = (a - \mu(\theta))^2$$

2. **absolute error** loss,

$$L(a, \theta) = |a - \mu(\theta)|$$

Loss functions in testing

- goal: decide whether $H_0 : \theta \in \Theta_0$ is true
- decision $a \in \{0, 1\}$ (true / not true)

Possible loss function:

$$L(a, \theta) = \begin{cases} 1 & \text{if } a = 1,\ \theta \in \Theta_0 \\ c & \text{if } a = 0,\ \theta \notin \Theta_0 \\ 0 & \text{else.} \end{cases}$$

	truth	
decision a	$\theta \in \Theta_0$	$\theta \notin \Theta_0$
0	0	c
1	1	0

Risk function

$$R(\delta,\theta) = E_\theta[L(\delta(X),\theta)].$$

- expected loss of a decision function δ
- R is a function of the true state of the world θ
- crucial intermediate object in evaluating a decision function
- small $R \Leftrightarrow$ good δ
- δ might be good for some θ, bad for other θ
- decision theory deals with this trade-off

Example: estimation of mean

- observe $X \sim N(\mu, 1)$
- want to estimate μ
- $L(a, \theta) = (a - \mu(\theta))^2$
- $\delta(X) = \alpha + \beta \cdot X$

Practice problem (Estimation of means)

Find the risk function for this decision problem.

Variance / Bias trade-off

Solution:

$$\begin{aligned} R(\delta,\mu) &= E[(\delta(X)-\mu)^2] \\ &= \mathrm{Var}(\delta(X)) + \mathrm{Bias}(\delta(X))^2 \\ &= \beta^2\,\mathrm{Var}(X) + (\alpha+\beta E[X]-E[X])^2 \\ &= \beta^2 + (\alpha+(\beta-1)\mu)^2. \end{aligned}$$

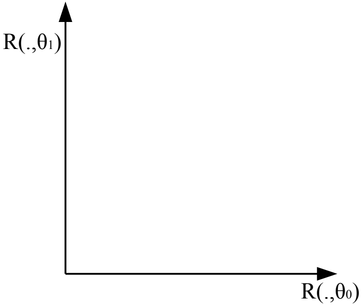
- equality 1 and 2: always true for squared error loss
- Choosing *b* (and *a*) involves a trade-off of bias and variance,
- this trade-off depends on μ.

Optimality criteria

- ranking provided by the risk function is multidimensional:
- a ranking of performance between decision functions for every θ
- to get a global comparison of their performance, have to aggregate this ranking into a global ranking
- preference relationship on space of risk functions
  $\Rightarrow$ preference relationship on space of decision functions

Illustrations for intuition

- suppose θ can only take two values,
- $\Rightarrow$ risk functions are points in a 2D-graph,
- each axis corresponds to $R(\delta, \theta)$ for $\theta = \theta_0, \theta_1$



Three approaches to get a global ranking

1. **partial ordering**:
   a decision function is better relative to another
   if it is better for *every* θ
2. complete ordering, **weighted average**:
   a decision function is better relative to another
   if a weighted average of risk across θ is lower
   weights $\sim$ prior distribution
3. complete ordering, **worst case**:
   a decision function is better relative to another
   if it is better under its worst-case scenario.

Approach 1: Admissibility

Dominance:
δ is said to dominate another function δ' if

$$R(\delta, \theta) \leq R(\delta', \theta)$$

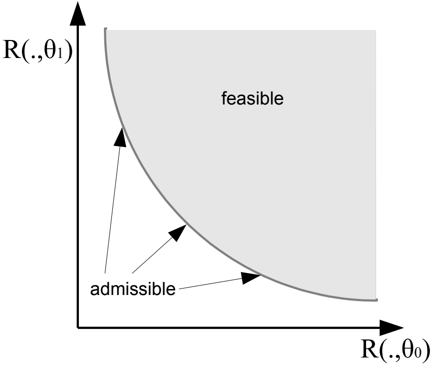
for all θ, and

$$R(\delta, \theta) < R(\delta', \theta)$$

for at least one θ.

Admissibility:
decisions functions which are not dominated are called admissible, all other decision functions are inadmissible.

Figure: Feasible and admissible risk functions



- admissibility $\sim$ “Pareto frontier”
- dominance only generates a partial ordering of decision functions
- in general: many different admissible decision functions.

Practice problem

- you observe $X_i \sim^{iid} N(\mu, 1)$, $i = 1, \ldots, n$
- your goal is to estimate μ, with squared error loss
- consider the estimators
  1. $\delta(X) = X_1$
  2. $\delta(X) = \frac{1}{n} \sum_i X_i$
- can you show that one of them is inadmissible?

Approach 2: Bayes optimality

- natural approach for economists:
- trade off risk across different θ
- by assigning weights $\pi(\theta)$ to each θ

Integrated risk:

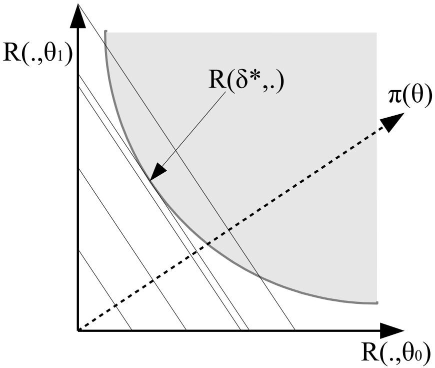
$$R(\delta,\pi) = \int R(\delta,\theta)\pi(\theta)d\theta.$$

Bayes decision function:
minimizes integrated risk,

$$\delta^* = \operatorname*{argmin}_{\delta} R(\delta, \pi).$$

- Integrated risk $\sim$ linear indifference planes in space of risk functions
- prior $\sim$ normal vector for indifference planes

Figure: Bayes optimality



Decision weights as prior probabilities

- suppose $0 < \int \pi(\theta) d\theta < \infty$
- then wlog $\int \pi(\theta) d\theta = 1$ (normalize)
- if additionally $\pi \geq 0$
- then π is called a prior distribution

Posterior

- suppose π is a prior distribution
- **posterior distribution:**

$$\pi(\theta|X) = \frac{f(X|\theta)\pi(\theta)}{m(X)}$$

- normalizing constant $=$ prior likelihood of X

$$m(X) = \int f(X|\theta)\pi(\theta)d\theta$$

Practice problem

- you observe $X \sim N(\theta, 1)$
- consider the prior
$$\theta \sim N(0, \tau^2)$$
- calculate
  1. $\pi(\theta|X)$
  2. $m(X)$

Posterior expected loss

$$R(\delta, \pi|X) := \int L(\delta(X), \theta)\pi(\theta|X)d\theta$$

Proposition

Any Bayes decision function δ^*
can be obtained by minimizing $R(\delta, \pi|X)$
through choice of $\delta(X)$ for every X.

Practice problem

Show that this is true.

Hint: show first that

$$R(\delta, \pi) = \int R(\delta(X), \pi|X)m(X)dX.$$

Bayes estimator with quadratic loss

- assume quadratic loss, $L(a,\theta) = (a - \mu(\theta))^2$
- posterior expected loss:

$$\begin{aligned} R(\delta,\pi|X) &= E_{\theta|X}\left[L(\delta(X),\theta)|X\right] \\ &= E_{\theta|X}\left[(\delta(X) - \mu(\theta))^2|X\right] \\ &= \mathrm{Var}(\mu(\theta)|X) + (\delta(X) - E[\mu(\theta)|X])^2 \end{aligned}$$

- Bayes estimator minimizes posterior expected loss $\Rightarrow$

$$\delta^*(X) = E[\mu(\theta)|X].$$

Practice problem

- you observe $X \sim N(\theta, 1)$
- your goal is to estimate θ, with squared error loss
- consider the prior

$$\theta \sim N(0, \tau^2)$$

- calculate
  1. $R(\delta(X), \pi|X)$
  2. $R(\delta, \pi)$
  3. the Bayes optimal estimator δ^*

Practice problem

- you observe X_i iid., $X_i \in \{1, 2, \ldots, k\}$, $P(X_i = j) = \theta_j$
- consider the so called Dirichlet prior, for $\alpha_j > 0$:

$$\pi(\theta) = const. \cdot \prod_{j=1}^{k} \theta_j^{\alpha_j - 1}$$

- calculate $\pi(\theta|X)$
- look up the Dirichlet distribution on Wikipedia
- calculate $E[\theta|X]$

Approach 3: Minimaxity

- Don't want to pick a prior?
- Can instead always assume the worst.
- worst $=\theta$ which maximizes risk

worst-case risk:

$$\overline{R}(\delta) = \sup_{\theta} R(\delta, \theta).$$

minimax decision function:

$$\delta^* = \operatorname*{argmin}_{\delta} \overline{R}(\delta) = \operatorname*{argmin}_{\delta} \sup_{\theta} R(\delta, \theta).$$

(does not always exist!)

Figure: Minimaxity ("Leontieff" indifference curves)