Theory of Statistical Inference - Lecture IV.1
STA422 and STA2162

Michael Evans
University of Toronto
https://utstat.utoronto.ca/mikevans/sta422/sta4222026.html

2026

- the previous likelihood-based methods discussed in the course can be thought of as part of the evidential approach, while decision theory constitutes the behavioral approach to developing a theory of statistics

- various ingredients are added to the basic inference base $I = (\{f_\theta : \theta \in \Theta\}, x)$ but one key addition is the loss function *Loss*

- as usual suppose we are interested in object $\psi = \Psi(\theta)$

Definition IV.1.1 Given a model $\{f_\theta : \theta \in \Theta\}$ and object of interest $\psi \in \Psi(\Theta)$ a loss function is a function

$$Loss : \Theta \times \Psi(\Theta) \to [0, \infty)$$

satisfying $Loss(\theta, \psi) = 0$ iff $\psi = \Psi(\theta)$ for all $\theta \in \Theta$. The set $\Psi(\Theta)$ is called the *action space*. ■

- the decision theory inference base is $I^{dec} = (\{f_\theta : \theta \in \Theta\}, Loss, x)$

- so if θ is the true value, namely, f_θ is the distribution that produced the data x, the true value of ψ is $\Psi(\theta)$ and no loss (penalty) is incurred if we "decide" that ψ takes the value $\Psi(\theta)$ and otherwise a loss is incurred

- loss can be thought of as the negative of a utility function as these arise in various fields but there is no notion of achieving a gain in statistical theory

- some approaches to decision theory will allow for utility rather than loss so then the goal is to maximize utility rather than minimize loss

- an axiomatic development of decision theory based on utility can be found in Savage (1954) The Foundations of Statistics

- see a list of the 7 axioms together with some discussion in Evans (2015) Measuring Statistical Evidence Using Relative Belief (in particular, the Allais paradox and the sure thing principle which is one of the axioms)

- what are some commonly used loss functions?

- L^2 or *quadratic loss*: when $\psi \in \mathbb{R}^k$, then

$$Loss(\theta, \psi) = (\psi - \Psi(\theta))^t A(\psi - \Psi(\theta))$$

for some fixed positive definite $A \in \mathbb{R}^{k \times k}$ (often the identity matrix)

- L^1 or *absolute error loss*: when $\psi \in \mathbb{R}^k$, then

$$Loss(\theta, \psi) = \sum_{i=1}^{k} w_i |\psi_i - \Psi_i(\theta)|$$

for some fixed $w_i > 0$ (often $w_i = 1$)

- *0-1 loss*: $Loss(\theta, \psi) = 1 - I_{\{\Psi(\theta)\}}(\psi)$

- is there a "correct" loss?

- is a loss function a falsifiable component of the inference base? in general, it doesn't seem so

- there are *intrinsic loss functions* formed from the model, or prior in Bayesian contexts, so these are falsifiable via model checking and checking for prior-data conflict, respectively (project?)

- the idea in decision theory is that, based on the data x, we want to choose a value $\psi \in \Psi(\Theta)$ that in some sense minimizes losses

- first we formalize the choice mechanism via a decision function

Definition IV.1.2 Given $I^{dec} = (\{f_\theta : \theta \in \Theta\}, Loss, x)$ a *decision function* δ is such that for each $x \in \mathcal{X}$, then $\delta(x, \cdot)$ is a probability measure on $\Psi(\Theta)$. If $\delta(x, d(x)) = 1$ for each x and some function $d : \mathcal{X} \to \Psi(\Theta)$, then it is a *nonrandomized decision function* and otherwise a *randomized decision function*. ■

- the idea is that after observing x, a decision is generated via $\psi \sim \delta(x, \cdot)$

- in nonconvex problems randomized decisions are necessary

- let $\mathcal{D}(I^{dec})$ denote the set of all decision functions for I^{dec} and $D(I^{dec})$ denote the set of all nonrandomized decision functions for I^{dec}

Lemma IV.1.1 $\mathcal{D}(I^{dec})$ is convex.
Proof: For $\alpha \in [0,1], \delta_1, \delta_2 \in \mathcal{D}(I^{dec})$, then

$$\delta(x,\cdot) = \alpha\delta_1(x,\cdot) + (1-\alpha)\delta_2(x,\cdot)$$

is a probability measure on $\Psi(\Theta)$ for every x and so $\delta \in \mathcal{D}(I^{dec})$. ■

Lemma IV.1. 2 $D(I^{dec})$ is convex iff $\Psi(\Theta)$ is convex.
Proof: $\Longrightarrow$] If $\psi_1, \psi_2 \in \Psi(\Theta), \alpha \in [0,1]$ and $d_1(x) \equiv \psi_1, d_2(x) \equiv \psi_2$, then $d_1, d_2 \in D(I^{dec})$ and so $\alpha d_1 + (1-\alpha)d_2 \in D(I^{dec})$ implies $\alpha\psi_1 + (1-\alpha)\psi_2 \in \Psi(\Theta)$.
$\Longleftarrow$] For any $\psi_1, \psi_2 \in \Psi(\Theta)$, then $\alpha\psi_1 + (1-\alpha)\psi_2 \in \Psi(\Theta)$ which implies $\alpha d_1 + (1-\alpha)d_2 \in D(I^{dec})$ for any $d_1, d_2 \in D(I^{dec})$. ■

- how to choose $\delta \in \mathcal{D}(I^{dec})$?

- note for each θ and decision function δ we have $x \sim f_\theta$ and $\psi \mid x \sim \delta(x, \cdot)$ and this induces a probability distribution for $L(\theta, \psi)$

- we want this distribution to concentrate as closely as possible about 0 for every $\theta \in \Theta$

- how to measure this? we could choose a quantile (say median) of the distribution but in decision theory the (arbitrary?) choice is the mean

Definition IV.1.3 Given $I^{dec} = (\{f_\theta : \theta \in \Theta\}, Loss, x)$ then the *risk function* of decision function δ is the function $R : \Theta \times \mathcal{D}(I^{dec})$ given by

$$R(\theta, \delta) = E_\theta(E_{\delta(x,\cdot)}(L(\theta, \psi))) = \int_{\mathcal{X}} \int_{\Psi(\Theta)} Loss(\theta, \psi)\, \delta(x, d\psi)\, P_\theta(dx). \blacksquare$$

- when δ is nonrandomized this becomes

$$R(\theta, d) = E_\theta(L(\theta, d)) = \int_{\mathcal{X}} Loss(\theta, d(x))\, P_\theta(dx)$$

- so a solution to the decision problem is a $\delta_{opt} \in \mathcal{D}(I^{dec})$ that satisfies

$$R(\theta, \delta_{opt}) = \inf_{\delta \in \mathcal{D}(I^{dec})} R(\theta, \delta)$$

for every $\theta \in \Theta$

Example IV.1.1 *Hypothesis testing*

- $\Theta = H_0 \cup H_a, H_0 \cap H_a = \phi$ the *null* versus the *alternative*

- $\psi \in \Psi(\Theta) = \{H_0, H_a\}$ and note that $\Psi(\Theta)$ is not convex

- the aim is to identify where θ_{true} lies either in H_0 or H_a

- using 0-1 loss

$$Loss(\theta, \psi) = \begin{cases} 0 & \text{when } \theta \in H_0, \psi = H_0 \text{ or } \theta \in H_a, \psi = H_a \\ 1 & \text{otherwise} \end{cases}$$

- there are two errors:

  type I error $= \theta \in H_0, \psi = H_a$ *or "*H_0 *is rejected when it is true"*
  type II error $= \theta \in H_a, \psi = H_0$ *or "*H_0 *is accepted when it is false"*

- note - the decision is to either accept H_0 as true or reject H_0 as false, there is no underlying idea concerning evidence with an associated strength measure

- for a $\delta \in \mathcal{D}(I^{dec})$ then define $\varphi : \mathcal{X} \to [0,1]$ by $\varphi(x) = \delta(x, H_a) =$ the probability that H_0 is rejected when x is observed is called a *test function* for H_0 versus H_a

- note - $\delta(x, H_0) = 1 - \delta(x, H_a) = 1 - \varphi(x)$ so $\delta \leftrightarrow \varphi$ then

$$\begin{aligned} R(\theta, \delta) &= \int_{\mathcal{X}} \int_{\Psi(\Theta)} Loss(\theta, \psi)\, \delta(x, d\psi)\, P_\theta(dx) \\ &= \int_{\mathcal{X}} (I_{H_0}(\theta)\delta(x, H_a) + I_{H_a}(\theta)\delta(x, H_0))\, P_\theta(dx) \\ &= \int_{\mathcal{X}} (I_{H_0}(\theta)\varphi(x) + I_{H_a}(\theta)(1 - \varphi(x)))\, P_\theta(dx) \\ &= I_{H_0}(\theta) E_\theta(\varphi) + I_{H_a}(\theta)(1 - E_\theta(\varphi)) \end{aligned}$$

- an optimal φ will minimize $E_\theta(\varphi)$ for each $\theta \in H_0$ (minimizing type I error) and maximize $\beta(\theta) = E_\theta(\varphi)$ for each $\theta \in H_a$ (minimizing type II error by maximizing the *power function* β of the test) ■

Example IV.1. 2 *Estimation with convex loss*

- suppose $\Psi(\Theta) \subset \mathbb{R}^k$ is convex and for each θ and the loss function is convex, namely, for all $\alpha \in [0,1]$

$$L(\theta, \alpha\psi_1 + (1-\alpha)\psi_2) \leq \alpha L(\theta, \psi_1) + (1-\alpha)L(\theta, \psi_2)$$

Lemma IV.1.3 Both quadratic loss and absolute error loss are convex losses.

Proof: For absolute error loss

$$L(\theta, \alpha\psi_1 + (1-\alpha)\psi_2) = \sum_{i=1}^{k} |\alpha\psi_{1i} + (1-\alpha)\psi_{2i} - \Psi_i(\theta)|$$

$$= \sum_{i=1}^{k} |\alpha(\psi_{1i} - \Psi_i(\theta)) + (1-\alpha)(\psi_{2i} - \Psi_i(\theta))|$$

$$\overset{\triangle \text{ inequality}}{\leq} \sum_{i=1}^{k} (\alpha|\psi_{1i} - \Psi_i(\theta)| + (1-\alpha)|\psi_{2i} - \Psi_i(\theta)|)$$

$$= \alpha L(\theta, \psi_1) + (1-\alpha)L(\theta, \psi_2). \blacksquare$$

Exercise IV.1.1 Show quadratic loss is a convex loss function.

Lemma IV.1.3 In a decision problem with convex loss the search for an optimal δ can be restricted to $D(I^{dec})$, the nonrandomized decision functions provided either $d_\delta(x) = E_{\delta(x,\cdot)}(\psi)$ exists finitely or otherwise $R(\theta, \delta) = \infty$.
Proof: For $\delta \in \mathcal{D}(I^{dec})$ where $d_\delta(x) = E_{\delta(x,\cdot)}(\psi)$ exists finitely, then $d_\delta(x) \in \Psi(\Theta)$, and we have

$$R(\theta, \delta) = \int_{\mathcal{X}} \int_{\Psi(\Theta)} Loss(\theta, \psi)\, \delta(x, d\psi)\, P_\theta(dx)$$

$$\overset{\text{Jensen's ineq}}{\geq} \int_{\mathcal{X}} \int_{\Psi(\Theta)} Loss(\theta, d_\delta(x))\, P_\theta(dx)$$

so we can consider $d_\delta(x)$ instead of δ. If $R(\theta, \delta) = \infty$, then put $d(x) \equiv c$ for some constant c and then $R(\theta, d) = Loss(\theta, c) < \infty$. Define $d^*(x) = E_{\delta(x,\cdot)}(\psi)$ when this exists finitely and $d^*(x) = c$ otherwise and then $R(\theta, d) \leq R(\theta, \delta)$ for all θ. ■

- for convex loss estimation problems a nonrandomized decision function d is called an *estimator*

- for quadratic loss

$$R(\theta, d) = E_\theta((d - \Psi(\theta))^t (d - \Psi(\theta)))$$

is called the *mean-squared error* of the estimator d

- note - if $E_\theta(d)$ is finite, then

$$\begin{aligned} & R(\theta, d) = E_\theta(((d - E_\theta(d)) + (E_\theta(d) - \Psi(\theta)))^t (\cdot)) \\ = \; & E_\theta((d - E_\theta(d))^t (\cdot)) + 2E_\theta((d - E_\theta(d))^t (E_\theta(d) - \Psi(\theta))) + \\ & (E_\theta(d) - \Psi(\theta))^t (\cdot) \\ = \; & tr\left\{E_\theta((d - E_\theta(d))(\cdot)^t)\right\} + (E_\theta(d) - \Psi(\theta))^t (\cdot) \\ = \; & tr\left\{Var_\theta(d)\right\} + bias_\theta^2 \end{aligned}$$

■

Example IV.1.3 *Classification - Estimation without convex loss*

- suppose $\Theta = \cup_{i=1}^{k} C_i$ where $C_i \cap C_j = \phi$ whenever $i \neq j$

- so $\Psi(\Theta) = \{C_1, \ldots, C_k\}$ and 0-1 loss

$$Loss(\theta, \psi) = \begin{cases} 0 & \text{when } \theta \in C_i, \psi = C_i \\ 1 & \text{otherwise} \end{cases}$$

- so based on data we want decide on which set C_i contains the true value of θ

- clearly just a generalization of hypothesis testing where $k = 2$, $C_1 = H_0$ and $C_2 = H_a$

- but also this an estimation problem where we want to estimate (classify x) $\psi_{true} = C_i$ s.t. $\theta_{true} \in C_i$ but the problem is not convex

- $\delta(x, \cdot)$ is a probability measure on $\{C_1, \ldots, C_k\}$ and

$$R(\theta, \delta) = \sum_{i=1}^{k} I_{C_i}(\theta) E_\theta(\delta(x, \{C_i\}^c)$$

■

- when does an optimal $\delta \in \mathcal{D}(I^{dec})$ exist?

Theorem IV.1.1 $\delta \in \mathcal{D}(I^{dec})$ is optimal iff $\delta(x, \cdot)$ is degenerate at $\Psi(\theta)$ with P_θ probability 1 for every θ.
Proof: Let $\delta_\theta \in \mathcal{D}(I^{dec})$ be defined by $\delta_\theta(x, \{\Psi(\theta)\}) = 1$ for each $\theta \in \Theta$. Now $R(\theta, \delta_\theta) = 0$ for every θ. Now an optimal δ must satisfy

$$0 \leq R(\theta, \delta) \leq R(\theta, \delta_\theta) = 0$$

for every θ so

$$0 = R(\theta, \delta) = \int_{\mathcal{X}} \int_{\Psi(\Theta)} Loss(\theta, \psi)\, \delta(x, d\psi)\, P_\theta(dx)$$

and since $\int_{\Psi(\Theta)} Loss(\theta, \psi)\, \delta(x, d\psi) \geq 0$, this holds iff $\int_{\Psi(\Theta)} Loss(\theta, \psi)\, \delta(x, d\psi) = 0$ with P_θ probability 1 for every θ, and since $Loss(\theta, \psi) \geq 0$ with equality iff $\psi = \Psi(\theta)$ so $\delta(x, \{\Psi(\theta)\}) = 1$ with P_θ probability 1 for every θ. ■

Corollary IV.1.1 If there exist $\theta_1, \theta_2 \in \Theta$ s.t. $\Psi(\theta_1) \neq \Psi(\theta_2)$ and $P_{\theta_1}, P_{\theta_2}$ are not concentrated on essentially disjoint sets with respect to $P_{\theta_1}, P_{\theta_2}$, then there does not exist an optimal δ.
Proof: Suppose optimal δ exists. Let $N_\theta = \{x : \delta(x, \cdot)$ is not degenerate at $\Psi(\theta)\}$. Then by the Theorem $P_\theta(N_\theta) = 0$ for every θ. If $x \notin N_{\theta_1} \cup N_{\theta_2}$, then $\delta(x, \cdot)$ is degenerate at $\Psi(\theta_1)$ and $\Psi(\theta_2)$ which is impossible so $\mathcal{X} = N_{\theta_1} \cup N_{\theta_2}$. Also $P_{\theta_1}(N_{\theta_1} \cap N_{\theta_2}) = P_{\theta_2}(N_{\theta_1} \cap N_{\theta_2}) = 0$ so $N_{\theta_1}, N_{\theta_2}$ are essentially disjoint with respect to $P_{\theta_1}, P_{\theta_2}$. Therefore P_{θ_1} is concentrated on N_{θ_2} and P_{θ_2} is concentrated on N_{θ_1} which is a contradiction to the hypothesis. Therefore, no optimal δ exists. ■

- essentially these results say that the only time an optimal δ exists is when $\mathcal{X} = \cup_{\theta \in \Theta} \mathcal{X}_\theta$ where the $\mathcal{X}_\theta$ are mutually disjoint, Ψ is constant on each $\mathcal{X}_\theta$ and each P_θ is concentrated on $\mathcal{X}_\theta$

- so when we observe x we know $\mathcal{X}_\theta$ and so we know $\Psi(\theta)$

- this is not a statistical problem