Theory of Statistical Inference - Lecture IV.6
STA422 and STA2162

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

2026

- recall that the preference ordering $\preceq$ on $\mathcal{D}(I^{dec})$ is only a preorder so yo can have $\delta_1 \not\preceq \delta_2$ and $\delta_2 \not\preceq \delta_1$

- the basic idea is to find a method of totally ordering the elements of $\mathcal{D}(I^{dec})$ in the hopes that this will lead to a δ preferred to all others (not necessarily unique)

- there are two such approaches in common usage: minimaxity and Bayesian decision theory

Definition IV.6.1 $\delta \in \mathcal{D}(I^{dec})$ is *minimax* if $\sup_\theta R(\theta, \delta) = \inf_{\delta' \in \mathcal{D}(I^{dec})} \sup_\theta R(\theta, \delta')$.

- this totally orders $\mathcal{D}(I^{dec})$

Example IV.6.1 *location normal*

- $x_1, \ldots, x_n$ iid $N_k(\mu, \sigma_0^2)$ where $\mu \in \mathbb{R}$ is unknown and σ_0^2 is known

- problem is to estimate μ using quadratic error

- so the problem is convex and we can restrict to nonrandomized d that depend on the data only through the mss $\bar{x}$

- let $\Pi = N(0, \tau^2)$ be a prior probability measure on $\mathbb{R}$, then a Bayes rule d minimizes $E_\Pi(R(\mu, d(\bar{x})))$

- as is well-known the Bayes rule for this problem is given by the posterior mean which is

$$d(\bar{x}) = (n/\sigma_0^2 + 1/\tau^2)^{-1} n\bar{x}/\sigma_0^2 = a(\tau^2)\bar{x}$$

and note $a(\tau^2) \to 1$ as $\tau^2 \to \infty$

- now

$$\begin{aligned} R(\mu, d(\bar{x})) &= E_\mu(a(\tau^2)\bar{x} - \mu)^2 \\ &= E_\mu(a(\tau^2)(\bar{x} - \mu) - (1 - a(\tau^2))\mu)^2 \\ &= a^2(\tau^2)\sigma_0^2/n + (1 - a(\tau^2)^2\mu^2 \text{ and so} \\ E_\Pi(R(\mu, d(\bar{x}))) &= a^2(\tau^2)\sigma_0^2/n + (1 - a(\tau^2)^2\tau^2 \to \sigma_0^2/n \text{ as } \tau^2 \to \infty \end{aligned}$$

- for any other estimate $d_* \in D(I^{dec})$

$$E_\Pi(R(\mu, d(\bar{x}))) \leq E_\Pi(R(\mu, d_*(\bar{x}))) < \sup_\mu R(\mu, d_*)$$

and since $R(\mu, \bar{x}) = \sigma_0^2/n$ this implies that

$$\sigma_0^2/n = \sup_\mu R(\mu, \bar{x}) < \sup_\mu R(\mu, d_*)$$

and so $\bar{x}$ is minimax ■

Example IV.6.2 *location-scale normal*

- $x_1, \ldots, x_n$ iid $N_k(\mu, \sigma_0^2)$ where $(\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$ unknown

- problem is to estimate μ using quadratic error

- so the problem is convex and we can restrict to nonrandomized d that depend on the data only through the mss $(\bar{x}, s^2)$ where $s^2 = \sum_{i=1}^n (x_i - \bar{x})^2$

- $R((\mu, \sigma^2), \bar{x}) = \sigma^2/n$ and for fixed σ^2 previous example implies

$$\frac{\sigma^2}{n} = \sup_{\mu} R((\mu, \sigma^2), \bar{x}) \leq \sup_{\mu} R((\mu, \sigma^2), d(\bar{x}))$$

for any other $d \in D(I^{dec})$

- now $\sup_{(\mu,\sigma^2)} R((\mu, \sigma^2), \bar{x}) = \infty$ which implies

$$\sup_{(\mu,\sigma^2)} R((\mu, \sigma^2), d(\bar{x})) \geq \sup_{\sigma^2} \sup_{\mu} R((\mu, \sigma^2), d(\bar{x})) = \infty$$

- therefore, all $d \in D(I^{dec})$ are minimax ■

- minimax regret as a possible fix to the overall minimax problem (?)

Definition IV.6.2 The *regret function* for $\delta \in \mathcal{D}(I^{dec})$ is given by

$$Regret(\theta, \delta) = R(\theta, \delta) - \inf_{\delta^* \in \mathcal{D}^*(I^{dec})} R(\theta, \delta^*)$$

where $\mathcal{D}^*(I^{dec}) \subset \mathcal{D}(I^{dec})$. A decision function δ_θ satisfying $R(\theta, \delta_\theta) = \inf_{\delta^* \in \mathcal{D}^*(I^{dec})} R(\theta, \delta^*)$ is called an *oracle decision function* and $\inf_{\delta^* \in \mathcal{D}^*(I^{dec})} R(\theta, \delta^*)$ is the *oracle risk*.$\inf_{\delta^* \in \mathcal{D}^*(I^{dec})} R(\theta, \delta^*)$. If

$$\sup_\theta Regret(\theta, \delta_0) = \inf_\delta \sup_\theta Regret(\theta, \delta)$$

then δ is a *minimax regret decision function*. ■

- note - if $\mathcal{D}^*(I^{dec}) = \mathcal{D}(I^{dec})$, then $\inf_{\delta^* \in \mathcal{D}^*(I^{dec})} R(\theta, \delta^*) = 0$ because, for fixed θ, then $\delta_\theta(x, \{\Psi(\theta)\}) = 1$ has $Loss(\theta, \Psi(\theta)) = 0$ which implies $R(\theta, \delta_\theta) = 0$ and so $Regret(\theta, \delta) = R(\theta, \delta)$ and minimax regret is just minimax and nothing is accomplished over minimaxity, so need to restrict $\mathcal{D}^*(I^{dec})$

Example IV.6.3 *location-scale normal (continued)*

- define oracle estimator to be $\bar{x}$ because it UMVU, optimal invariant and admissible which implies so oracle risk is σ^2/n (could take $\mathcal{D}^*(I^{dec}) =$ set of all unbiased estimators)

- this implies that $Regret(\theta, \bar{x}) = 0$ and admissibility implies that $\bar{x}$ is the minimax regret estimator

- now suppose it is known that $-M \leq \mu \leq M, \sigma^2_{\min} \leq \sigma^2 \leq \sigma^2_{\max}$

- let $d(x) \equiv 0$ so $R((\mu, \sigma^2), d) = \mu^2$ so

$$\begin{aligned} Regret((\mu, \sigma^2), d) &= \mu^2 - \sigma^2/n \text{ which implies} \\ \sup_{(\mu,\sigma^2)} Regret((\mu, \sigma^2), d) &= M^2 - \sigma^2_{\min}/n < \sigma^2/n \end{aligned}$$

when $M < \sqrt{(\sigma^2 + \sigma^2_{\min})/n}$ and so $\bar{x}$ would no longer be minimax regret although it would be asymptotically ■

- we add a component to $I^{Bayes\ dec} = (\{f_\theta : \theta \in \Theta\}, Loss, \pi, x))$ where π is a probability density on Θ called the *prior*

Definition IV.6.2 For I^{dec} the prior risk of $\delta \in \mathcal{D}(I^{dec})$ is $r(\delta) = E_\pi(R(\theta, \delta))$ and δ_0 is a Bayes rule if $r(\delta_0) = \inf_{\delta \in \mathcal{D}(I^{dec})} r(\delta)$. ■

- typically a Bayes rule exists and so is a solution to the decision problem $I^{Bayes\ dec}$

- note that (θ, x) has a joint probability distribution (interpreting density f_θ as the conditional density of x given θ) with density $\pi(\theta) f_\theta(x)$

- when we observe x the principle of conditional probability says we replace the prior π by the *posterior density*

$$\pi(\theta \,|\, x) = \frac{\pi(\theta) f_\theta(x)}{m(x)} \text{ where } m(x) = \int_\Theta \pi(\theta) f_\theta(x)\, d\theta$$

is the prior density for the data called the *prior predictive*

- in the search for the Bayes rule the following result is useful

Proposition IV.6.1 If δ_0 minimizes the *posterior risk*

$$r(\delta \mid x) = \int_{\Theta} \int_{\Psi(\Theta)} Loss(\theta, \psi)\delta(x, d\psi)\pi(\theta \mid x)\, d\theta$$

for each $x \in \mathcal{X}$, then δ_0 is a Bayes rule.
Proof: For $\delta \in \mathcal{D}(I^{dec})$ we have

$$\begin{aligned} r(\delta) &= E_{\pi}(R(\theta, \delta)) = \int_{\Theta} R(\theta, \delta)\pi(\theta)\, d\theta \\ &= \int_{\Theta} \int_{\mathcal{X}} \int_{\Psi(\Theta)} Loss(\theta, \psi)\delta(x, d\psi) f_{\theta}(x)\pi(\theta)\, dx\, d\theta \\ &= \int_{\mathcal{X}} \left(\int_{\Theta} \int_{\Psi(\Theta)} Loss(\theta, \psi)\delta(x, d\psi)\pi(\theta \mid x)\, d\theta \right) m(x)\, dx \\ &= \int_{\mathcal{X}} r(\delta \mid x) m(x)\, dx \geq \int_{\mathcal{X}} r(\delta_0 \mid x) m(x)\, dx = r(\delta). \blacksquare \end{aligned}$$

- note - in general $L(\theta, \psi)$ is dependent on θ only through $\Psi(\theta)$ so can write $L(\theta, \psi) = L(\Psi(\theta), \psi)$

- and $\pi(\theta) = \pi(\theta \mid \psi)\pi_\Psi(\psi)$ where $\pi(\theta \mid \psi)$ is the density on the nuisance parameters and $\pi_\Psi(\psi)$ is the prior on the parameter of interest

- put $m_\psi(x) = \int_{\Psi^{-1}\{\psi\}} f_\theta(x)\pi(\theta \mid \psi)\, d\theta$ leading to model $\{m_\psi : \psi \in \Psi(\theta)\}$ and inference base for Ψ given by $I_\Psi^{Bayes\ dec} = (\{m_\psi : \psi \in \Psi(\theta)\}, Loss, \pi_\Psi, x)$ and

$$r(\delta) = \int_\Theta \int_{\mathcal{X}} \left(\int_{\Psi(\Theta)} Loss(\Psi(\theta), \psi)\delta(x, d\psi) \right) f_\theta(x)\pi(\theta)\, dx\, d\theta$$

$$= \int_{\Psi(\Theta)} \int_{\Psi^{-1}\{\psi'\}} \int_{\mathcal{X}} E_{\delta(x,\cdot)}(Loss(\psi', \psi)) f_\theta(x)\pi(\theta \mid \psi')\pi_\Psi(\psi')\, dx\, d\theta\, d\psi'$$

$$= \int_{\Psi(\Theta)} \int_{\mathcal{X}} E_{\delta(x,\cdot)}(Loss(\psi', \psi)) \int_{\Psi^{-1}\{\psi'\}} f_\theta(x)\pi(\theta \mid \psi)\, d\theta\, \pi_\Psi(\psi')\, dx\, d\psi'$$

$$= \int_{\Psi(\Theta)} \int_{\mathcal{X}} E_{\delta(x,\cdot)}(Loss(\psi', \psi)) m_{\psi'}(x)\, \pi_\Psi(\psi')\, dx\, d\psi'$$

- we conclude that the original decision problem

$$I^{Bayes\ dec} = (\{f_\theta : \theta \in \Theta\}, Loss, \pi, x)$$

can be replaced by

$$I_\Psi^{Bayes\ dec} = (\{m_\psi : \psi \in \Psi(\theta)\}, Loss, \pi_\Psi, x)$$

- note too

$$\begin{aligned}\pi_\Psi(\psi \mid x) &= \int_{\Psi^{-1}\{\psi\}} \pi(\theta \mid x)\, d\theta = \int_{\Psi^{-1}\{\psi\}} \frac{\pi(\theta) f_\theta(x)}{m(x)}\, d\theta \\ &= \frac{\pi_\Psi(\psi)}{m(x)} \int_{\Psi^{-1}\{\psi\}} \pi(\theta \mid \psi) f_\theta(x)\, d\theta = \frac{\pi_\Psi(\psi) m_\psi(x)}{m(x)}\end{aligned}$$

and

$$\begin{aligned}m(x) &= \int_\Theta f_\theta(x)\pi(\theta)\, d\theta = \int_{\Psi(\Theta)} \int_{\Psi^{-1}\{\psi\}} f_\theta(x)\, d\theta\, \pi_\Psi(\psi)\, d\psi \\ &= \int_{\Psi(\Theta)} m_\psi(x)\, \pi_\Psi(\psi)\, d\psi\end{aligned}$$

- we integrate out nuisance parameters, i.e., nuisance parameters aren't a problem in a proper prior Bayesian formulation

Example IV.6.4 *Estimation with quadratic loss*

- convexity implies we can restrict to nonrandomized rules d then assuming $E_{\pi_\Psi(\cdot\,|\,x)}(\psi)$ is finite

$$\begin{aligned} r(d\,|\,x) &= E_{\pi_\Psi(\cdot\,|\,x)}((d(x)-\psi)^t A(d-\psi)) \\ &= E_{\pi_\Psi(\cdot\,|\,x)}((d(x)-E_{\pi_\Psi(\cdot\,|\,x)}(\psi)+E_{\pi_\Psi(\cdot\,|\,x)}(\psi)-\psi)^t A(\cdot)) \\ &= (d(x)-E_{\pi_\Psi(\cdot\,|\,x)}(\psi))^t A(\cdot)+E_{\pi_\Psi(\cdot\,|\,x)}((E_{\pi_\Psi(\cdot\,|\,x)}(\psi)-\psi)^t A(\cdot)) \end{aligned}$$

and since each term is nonnegative this is minimized by taking $d(x)=E_{\pi_\Psi(\cdot\,|\,x)}(\psi)$ so the posterior mean is a Bayes rule ■

Example IV.6.5 *Hypothesis testing*

- $\Theta = H_0 \cup H_a, H_0 \cap H_a = \phi$

- recall with 0-1 loss and $\varphi(x) = \delta(x, H_a)$

$$\begin{aligned} r(\varphi \,|\, x) &= \int_{\Theta} \int_{\Psi(\Theta)} Loss(\theta, \psi) \delta(x, d\psi) \pi(\theta \,|\, x)\, d\theta \\ &= \int_{H_0} \varphi(x) \pi(\theta \,|\, x)\, d\theta + \int_{Ha} (1 - \varphi(x))) \pi(\theta \,|\, x)\, d\theta \\ &= \Pi(H_0 \,|\, x) \varphi(x) + \Pi(H_a \,|\, x)(1 - \varphi(x)) \end{aligned}$$

and so Bayes rule is given by

$$\varphi(x) = \begin{cases} 1 & \Pi(H_a \,|\, x) \geq \Pi(H_0 \,|\, x) \\ 0 & \Pi(H_a \,|\, x) < \Pi(H_0 \,|\, x) \end{cases}$$

- when $\Pi(H_a \,|\, x) = \Pi(H_0 \,|\, x) = 1/2$, it doesn't matter which you accept

- when $\Pi(H_0) = 0$ then $\Pi(H_0 \,|\, x) = 0$ and we always reject H_0 which doesn't make sense ■

- there are a variety of results that show that Bayes rules are admissible

Proposition IV.6.2 If δ is a Bayes rule and it is unique up to risk equivalence (two Bayes rules have the same risk functions) then δ is admissible.
Proof: Suppose $R(\theta, \delta_0) \leq R(\theta, \delta)$ for every θ. Then $r(\delta_0) \leq r(\delta)$ which implies $r(\delta_0) = r(\delta)$ which proves $R(\theta, \delta_0) = R(\theta, \delta)$. ■

- for any Bayes rule δ, if $R(\theta, \delta_0) \leq R(\theta, \delta)$ for every θ, then the above shows that $R(\theta, \delta_0) = R(\theta, \delta)$ with prior probability 1

Corollary IV.6.1 If $\Theta = \{\theta_1, \ldots, \theta_k\}$ and $\pi(\theta) > 0$ for all θ, then a Bayes rule is admissible.

Proposition IV.6.3 If $\Theta = \{\theta_1, \ldots, \theta_k\}$ and δ is admissible, then δ is a Bayes rule for some prior.
Proof: accept.

- there are a wide variety of results generalizing these results on admissibility

Personal Opinion

The only approach to decision theory that seems fully satisfactory is Bayesian decision theory. Still Bayesian decision theory requires a loss function, which is generally not falsifiable, so it seems unsatisfactory for scientific applications. More significant, however, is that Bayesian decision theory contains no measure of the reliability of the inferences, is based on the somewhat arbitrary choice of averaging losses and there is no clear definition of what is meant by statistical evidence. There are also examples where Bayes rules, like posterior means, seem less than satisfactory as does hypothesis testing when $\Pi(H_0) = 0$ *or even when it is small.*