Theory of Statistical Inference - Lecture IV.2
STA422 and STA2162

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

2026

- the theory fails to produce an optimal $\delta \in \mathcal{D}(I^{dec})$

- what to do?

- there are several strategies

(1) **Restriction Principles**
replace $\mathcal{D}(I^{dec})$ by $\mathcal{D}_*(I^{dec}) \subset \mathcal{D}(I^{dec})$ where all the elements possess some property and then search for an optimal $\delta \in \mathcal{D}_*(I^{dec})$; two properties of interest are *unbiasedness* and *invariance* under symmetry groups

(2) **Modify the Optimality Criterion**: retain $\mathcal{D}(I^{dec})$ but modify (weaken) the definition of optimality so that an optimal $\delta \in \mathcal{D}(I^{dec})$ exists; two such modifications are *minimaxity* and *Bayes rules*

- one restriction is to restrict to decision functions that depend on the data only through the value of a minimal sufficient statistic

Proposition IV.2.1 If T is sufficient for I^{dec}, then for $\delta \in \mathcal{D}(I^{dec})$ define δ_T by, when x is observed, then generate $z \sim P(\cdot \mid T)(T(x)$ and $\psi \sim \delta(z, \cdot)$. Then $\delta_T(x_1, \cdot) = \delta_T(x_2, \cdot)$ when $T(x_1) = T(x_2)$ and $R(\theta, \delta_T) = R(\theta, \delta)$ for every $\theta \in \Theta$. So, δ_T only depends on the data through the value of T and we can write $\delta_T(t, \cdot)$ for $t \in \mathcal{T} = T(\mathcal{X})$ instead.

Proof: For $B \subset \Psi(\Theta)$ we have

$$\begin{aligned} \delta_T(x_1, A) &= \int_{\mathcal{X}} \delta(z, A)\, P(dz \mid T)(T(x_1)) \\ &= \int_{\mathcal{X}} \delta(z, A)\, P(dz \mid T)(T(x_2)) = \delta_T(x_2, A). \end{aligned}$$

Also,

$$
\begin{aligned}
R(\theta,\delta) &= \int_{\mathcal{X}}\int_{\Psi(\Theta)} L(\theta,\psi)\,\delta(x,d\psi)\,P_\theta(dx) \\
&= \int_{\mathcal{T}}\int_{\mathcal{X}}\int_{\Psi(\Theta)} L(\theta,\psi)\,\delta(z,d\psi)\,P(dz\,|\,T)(t)\,P_{\theta T}(dt) \\
&= \int_{\mathcal{T}}\int_{\Psi(\Theta)} L(\theta,\psi)\,\left(\int_{\mathcal{X}}\delta(z,d\psi)\,P(dz\,|\,T)(t)\right)\,P_{\theta T}(dt) \\
&= \int_{\mathcal{T}}\int_{\Psi(\Theta)} L(\theta,\psi)\,\delta_T(t,d\psi)\,P_{\theta T}(dt) = R(\theta,\delta_T). \blacksquare
\end{aligned}
$$

- note that δ may be nonrandomized but now δ_T is randomized and the general restriction to decision functions that depend on the data only through the value of a minimal sufficient statistic requires this (controversial?)

- so, for sufficient T, we can restrict to

$$\mathcal{D}_T(I^{dec}) = \{\delta \in \mathcal{D}(I^{dec}) : \delta \text{ depends on the data only through } T\}$$

or replace $I^{dec} = (\{f_\theta : \theta \in \Theta\}, L, x)$ by $I_T^{dec} = (\{f_{\theta T} : \theta \in \Theta\}, L, T(x))$

- one restriction is to $D(I^{dec}) \subset \mathcal{D}(I^{dec})$ when I^{dec} is convex because the nonrandomized $d_\delta(x) = E_{\delta(x,\cdot)}(\psi)$ has risk no greater than δ, how does this interact with sufficiency?

Theorem IV.2.1 (*Rao-Blackwell*) If I^{dec} is convex and T is sufficient for I^{dec} then for $d \in D(I^{dec})$ where

$$d_T(x) = E_{P(dz \mid T)(T(x))}(d) = \int_{\mathcal{X}} d(z)\, P(dz \mid T)(T(x))$$

exists, then $d_T(x_1) = d_T(x_2)$ when $T(x_1) = T(x_2)$ and $R(\theta, d_T) \le R(\theta, d)$ for every $\theta \in \Theta$.

Proof: Again we can write d_T as $d_T(t)$, namely, as a function of $t \in \mathcal{T}$ and then

$$R(\theta, d) = \int_{\mathcal{T}} \int_{\mathcal{X}} L(\theta, d(z))\, P(dz \mid T)(t)\, P_{\theta T}(dt)$$

$$\overset{\text{Jensen ineq}}{\ge} \int_{\mathcal{T}} L(\theta, d_T(t))\, P_{\theta T}(dt) = R(\theta, d_T). \blacksquare$$

- *Rao-Blackwellize* an estimator: when appropriate replace d by d_T

- the restriction via sufficiency or nonrandomized rules do not guarantee an optimal member but they do no damage

- whatever strategy is employed the concept of admissibility is relevant

- this is based on a preorder (reflexive and transitive) on $\mathcal{D}(I^{dec}) : \delta_1$ is *preferred to* δ_2, written $\delta_1 \preceq \delta_2$ if $R(\theta, \delta_1) \le R(\theta, \delta_2)$ for every $\theta \in \Theta$ and δ_1 is *strictly preferred to* δ_2, written $\delta_1 \prec \delta_2$, if $\delta_1 \preceq \delta_2$ and there exists $\theta \in \Theta$ s.t. $R(\theta, \delta_1) < R(\theta, \delta_2)$

- note - it is possible that $\delta_1 \preceq \delta_2$ and $\delta_2 \preceq \delta_1$ but $\delta_1 \neq \delta_2$ and also $\preceq$ is not a total order

Definition IV.2.1 The decision function $\delta \in \mathcal{D}(I^{dec})$ is *admissible* if there doesn't exist δ_* such that $\delta_* \prec \delta$. Let $\mathcal{A}(I^{dec}) \subset \mathcal{D}(I^{dec})$ denote the set of admissible rules.

- note - admissibility is not necessarily a positive attribute but inadmissibility is to be considered, from the viewpoint of decision theory, as a strong negative

- one is not supposed to use inadmissible procedures because there is something that will do better with respect to risk

- one possible solution to the decision problem is to simply state $\mathcal{A}(I^{dec})$ but ...

Lemma IV.2.1 If there exists $\theta_0 \in \Theta$ s.t. P_θ is absolutely continuous with respect to P_{θ_0}, then δ_{θ_0} ($\delta_{\theta_0}(x, \{\Psi(\theta_0)\}) = 1$ with P_{θ_0} probability 1) is admissible.
Proof: Suppose $\delta \preceq \delta_{\theta_0}$. Then $0 \leq R(\theta_0, \delta) \leq R(\theta_0, \delta_{\theta_0}) = 0$ and so $R(\theta_0, \delta) = 0$, but this only happens if $\delta(x, \{\Psi(\theta_0)\}) = 1$ with P_{θ_0} probability 1. If $B = \{x : \delta(x, \cdot) \neq \delta_{\theta_0}(x, \cdot)\}$, then $P_{\theta_0}(B) = 0$ which implies that $P_\theta(B) = 0$ for every θ and so $\delta = \delta_{\theta_0}$. ■

- so for most of the commonly used statistical models, constants are admissible procedures and so this doesn't seem like a good solution as we need to somehow eliminate the "bad" admissible rules