ST5215 Advanced Statistical Theory, Lecture 11

HUANG Dongming

National University of Singapore

15 Sep 2020

Overview

Last time

- Statistical Decision Theory
- Statistical Inference

Today

- More on risk of estimators
- Minimax rules and Bayes rules
- UMVUE

Recap: Decision theory

Let X be a sample from a population $P \in \mathcal{P}$.

- A statistical decision is an action that we take after we observe X
- The set of allowable actions $\mathbb{A}$, endowed with a σ-field $\mathcal{F}_{\mathbb{A}}$
- A *decision rule* T: a measurable function from $(\mathcal{X}, \mathcal{F}_{\mathcal{X}})$ to $(\mathbb{A}, \mathcal{F}_{\mathbb{A}})$. If T is chosen, then we take the action $T(X) \in \mathbb{A}$ whence X is observed
- *Loss function*: a function $L : \mathcal{P} \times \mathbb{A} \to [0, \infty)$ that is Borel for each fixed $P \in \mathcal{P}$. If $X = x$ is observed and a decision rule T is chosen, then the "loss" in making a decision is $L(P, T(x))$
- *Risk*: the average loss $R_T(P) = E_P\left[L(P, T(X))\right]$
- T_1 is as good as T_2: if $R_{T_1}(P) \leq R_{T_2}(P)$ for all $P \in \mathcal{P}$
- T_1 is better than T_2: if T_1 is as good as T_2 and $R_{T_1}(P) < R_{T_2}(P)$ for some $P \in \mathcal{P}$
- T_* is optimal: if T_* is as good as any other rule
- T is admissible: if no rule is better than T

Example (1)

Suppose the parameter space is $\Theta = \{\theta_A, \theta_B, \theta_C\}$ and the rules are $\mathfrak{J} = \{T_1, T_2, T_3, T_4\}$.
The risk for each rule under each population is showed in the table.
Which rule is $\mathfrak{J}$-optimal? Are they $\mathfrak{J}$-admissible?

Rule $\setminus$ θ	θ_A	θ_B	θ_C	Optimal?	Admissible?
T_1	0	2	3	No	No
T_2	1	1	2	No	No
T_3	1	2	2	No	No
T_4	0	1	2	Yes	Yes

Now, suppose we do not consider T_4

Example (2)

Suppose the parameter space is $\Theta = \{\theta_A, \theta_B, \theta_C\}$ and the rules are $\mathfrak{J} = \{T_1, T_2, T_3\}$.
The risk for each rule under each population is showed in the table.
Which rule is $\mathfrak{J}$-optimal? Are they $\mathfrak{J}$-admissible?

Rule $\backslash$ θ	θ_A	θ_B	θ_C	Optimal?	Admissible?
T_1	0	2	3	No	Yes
T_2	1	1	2	No	Yes
T_3	1	2	3	No	No

Remark. Although T_2 is not optimal, it is **minimax**

Exercise on Optimality and Admissibility

Suppose $X_1, \ldots, X_n$ are i.i.d. with mean θ and variance σ^2.
Assume σ is known and $\theta \in [0, \infty)$.
Consider $\mathfrak{J}$ the class of estimators $\hat{\theta}_c = c\bar{X}$ for $c \in [0, 1]$.

- Last week, we know that the MSE of $\hat{\theta}_c$ is

$$R_c(\theta) = (1-c)^2\theta^2 + c^2\sigma^2/n$$

In $\mathfrak{J}$, which estimator is optimal? Which is admissible?

- For each θ, $c^*_\theta = \theta^2 / \left(\theta^2 + \sigma^2/n\right)$ is the unique minimizer of $R_c(\theta)$, so there is no $\mathfrak{J}$-optimal rule
- For each $c \in [0, 1)$, we can find $\theta^*_c = \sigma\sqrt{\frac{c}{n(1-c)}}$ such that c is the unique minimizer of $R_c(\theta^*_c)$. Therefore, $\hat{\theta}_c$ is $\mathfrak{J}$-admissible for each $c \in [0, 1)$
- $\hat{\theta}_1$ is also $\mathfrak{J}$-admissible: to compare with any $\hat{\theta}_c$, where $c \in [0, 1)$, look at $R_c(\theta)/R_1(\theta) = n(1-c)^2\theta^2/\sigma^2 + c^2$, which must be greater than 1 for θ large enough

Minimaxity

- The risk $R_T(P)$ is defined for a given $P \in \mathcal{P}$
- A decision rule that works well for one population may work extremely bad for another
- To find a good decision rule, we can consider some characteristic R_T of $R_T(P)$ for a given decision rule T, and then minimize R_T over $T \in \mathfrak{J}$
- One useful way is to consider the worst risk

Definition

Let $\mathfrak{J}$ be a class of decision rules. A decision rule $T_* \in \mathfrak{J}$ is called $\mathfrak{J}$-minimax if $\sup_{P \in \mathcal{P}} R_{T_*}(P) \leq \sup_{P \in \mathcal{P}} R_T(P)$ for any $T \in \mathfrak{J}$

Remark. In words, a minimax rule tries to do as well as possible in the worst case.

Example Revisited

Suppose $X_1, \dots, X_n$ are i.i.d. with mean θ and variance σ^2.
Assume σ is known and $\theta \in [0, \infty)$.
Consider $\mathfrak{J}$ the class of estimators $\hat{\theta}_c = c\bar{X}$ for $c \in [0, 1]$.

- The MSE of $\hat{\theta}_c$ is

$$R_c(\theta) = (1 - c)^2\theta^2 + c^2\sigma^2/n$$

- For each $c \in [0, 1)$, $\sup_{\theta \in [0,\infty)} R_c(\theta) = \infty$
- $\bar{X}$ is $\mathfrak{J}$-minimax: $\sup_{\theta \in [0,\infty)} R_1(\theta) = \sigma^2/n < \infty$

Bayes Rule

- It is also useful to consider an average of $R_T(P)$ over $P \in \mathcal{P}$:

$$r_T(\Pi) = \int_{\mathcal{P}} R_T(P) d\Pi(P),$$

  where Π is a known probability measure on $(\mathcal{P}, \mathcal{F}_{\mathcal{P}})$ with an appropriate σ-field $\mathcal{F}_{\mathcal{P}}$
- $r_T(\Pi)$ is called the *Bayes risk* of T w.r.t. Π
- For a parametric family $\{P_\theta : \theta \in \Theta\}$, we can simply use a probability measure on Θ and define

$$r_T(\pi) = \int_{\Theta} R_T(P_\theta) d\pi(\theta),$$

If $T_* \in \mathfrak{J}$ and $r_{T_*}(\Pi) \leq r_T(\Pi)$ for any $T \in \mathfrak{J}$, then T_* is called a $\mathfrak{J}$-*Bayes rule* (or *Bayes rule* when $\mathfrak{J}$ contains all possible rules) w.r.t. Π

Remark. A Bayes risk is just one summary of $R_T(P)$ over $P \in \mathcal{P}$. This notion does not rely on *Bayesian statistics*, in which parameters are viewed as unobserved random variables.

Example Revisited

Suppose $X_1, \ldots, X_n$ are i.i.d. with mean θ and variance σ^2.
Assume σ is known and $\theta \in [0, \infty)$.
Consider $\mathfrak{J}$ the class of estimators $\hat{\theta}_c = c\bar{X}$ for $c \in [0, 1]$.

- The MSE of $\hat{\theta}_c$ is

$$R_c(\theta) = (1-c)^2\theta^2 + c^2\sigma^2/n$$

- Now consider a probability on Θ : $\pi = \mathrm{Exp}(1)$ (the exponential distribution with rate 1)
- The Bayes risk of $\hat{\theta}_c$ is $r_\pi(\hat{\theta}_c) = 2(1-c)^2 + c^2\sigma^2/n$
- When $c = 2/(2+\sigma^2/n)$, $\hat{\theta}_c$ is the $\mathfrak{J}$-Bayes rule

Finding a Bayes rule

- We introduce two random elements $\tilde{\theta} \sim \pi$, and $X \mid \tilde{\theta} \sim P_{\tilde{\theta}}$
- Then the Bayes risk $r_\pi(T)$ can be expressed as $E\left[L(\tilde{\theta}, T(X))\right]$, where E is taken jointly over $(\tilde{\theta}, X)$
- Using the tower property of conditional expectation, we may rewrite the Bayes risk as

$$E\left\{E\left[L(\tilde{\theta}, T(X)) \mid X\right]\right\} \tag{1}$$

- If for every x, *the conditional risk* $E\left[L(\tilde{\theta}, a) \mid X = x\right]$ is minimized at $a = T_*(x)$, then T_* is the Bayes rule w.r.t. π

Example: Squared Error

- Consider the squared error for estimation $L(\theta, a) = (\theta - a)^2$ and some probability π on Θ
- After introducing $\tilde{\theta} \sim \pi$ and $X \mid \tilde{\theta} \sim P_{\tilde{\theta}}$, we need to minimize over a

$$E\left[(\tilde{\theta} - a)^2 \mid X = x\right]$$

- The Bayes estimator turns out to be the *posterior mean* $T(X) := E[\tilde{\theta} \mid X]$, where the expectation is taken with respect to the conditional distribution of $\tilde{\theta}$ given X (also known as the *posterior distribution*)

Unbiased Estimators

- Let X be a sample from an unknown population $P \in \mathcal{P}$ and θ be a real-valued parameter related to P
- Recall that an estimator $T(X)$ of θ is unbiased if and only if $E[T(X)] = \theta$ for any $P \in \mathcal{P}$
- If there exists an unbiased estimator of θ, then θ is called an *estimable* parameter
- For squared error loss, the risk of an unbiased estimator is equal to its variance
- We can compare unbiased estimators by their variance

UMVUE

Definition (UMVUE)

An unbiased estimator $T(X)$ of θ is called the *uniformly minimum variance unbiased estimator (UMVUE)* if and only if $\mathrm{Var}(T(X)) \leq \mathrm{Var}(U(X))$ for any $P \in \mathcal{P}$ and any other unbiased estimator $U(X)$ of θ.

- "Uniformly" refers to "for any $P \in \mathcal{P}$"
- A UMVUE estimator is $\mathfrak{J}$-optimal in MSE with $\mathfrak{J}$ being the class of all unbiased estimators
- Rao-Blackwell theorem implies that the variance of the conditional expectation of an unbiased estimator given a sufficient statistic is smaller
- In fact, if the sufficient statistic is also complete, this variance is uniformly minimal

Theorem (Lehmann-Scheffé)

Suppose that there exists a sufficient and complete statistic $T(X)$ for $P \in \mathcal{P}$, and θ is related to P. If θ is estimable, then there is a unique unbiased estimator of θ that is of the form $h(T)$ with a Borel function h. Furthermore, $h(T)$ is the unique UMVUE of θ.

Proof:

- By assumption, there is an unbiased estimator $\hat{\theta}$ for θ
- Let $h(T) = E(\hat{\theta} \mid T)$. The LHS does not depend on P because T is sufficient
- Then $Eh(T) = E\hat{\theta} = \theta$, i.e., $h(T)$ is unbiased for θ
- The squared error loss is convex. By Rao-Blackwell theorem, for any other unbiased estimator of θ, its conditional expectation given T, say, $g(T)$, is as good
- Since both $h(T)$ and $g(T)$ are unbiased, $E\{h(T) - g(T)\} = 0$, $\forall P \in \mathcal{P}$
- The completeness of T implies that $h - g = 0, \ \ \mathcal{P}$-a.s.
- Therefore, $h(T)$ is an UMVUE and is unique

Example: Uniform distributions

Let $X_1, ..., X_n$ be i.i.d. from the uniform distribution on $(0, \theta)$, $\theta > 0$. Find the UMVUE of θ.

- In previous lectures, we have shown that the order statistic $X_{(n)}$ is sufficient and complete with Lebesgue p.d.f. $n\theta^{-n}x^{n-1}I_{(0,\theta)}(x)$
- We observe that

$$E_\theta X_{(n)} = n\theta^{-n}\int_0^\theta x^n dx = \frac{n}{n+1}\theta. \tag{2}$$

- Therefore, $E_\theta\{(n+1)X_{(n)}/n\} = \theta$ for all $\theta > 0$
- By Lehmann-Scheffé theorem, $\hat{\theta} = (n+1)X_{(n)}/n$ is the unique UMVUE of θ

Plans ...

- Next lecture: methods of finding UMVUEs
- Preparation for the Online Midterm Exam

Problem 1 in Homework 1

If $f : \mathcal{R} \mapsto \mathcal{R}$ is a continuous function, then it is Borel measurable **Recall**

- Denote by $\mathcal{O}$ the collection of all open sets in $\mathcal{R}$
- The Borel σ-field $\mathcal{B}$ is the smallest σ-field that contains $\mathcal{O}$
- The inverse image is defined as $f^{-1}(A) = \{x : f(x) \in A\}$

Proof:

- Consider $\mathcal{F} = \{A \subset \mathcal{R} : f^{-1}(A) \in \mathcal{B}\}$
- We need to show $\mathcal{F}$ is a σ-field that contains $\mathcal{O}$
- Check
  1. $f^{-1}(\emptyset) = \emptyset \in \mathcal{B}$. So $\emptyset \in \mathcal{F}$
  2. If $A \in \mathcal{F}$, then $f^{-1}(A^c) = \left(f^{-1}(A)\right)^c$. So $A^c \in \mathcal{F}$
  3. If $A_i \in \mathcal{F}$ for all $i \in \mathbf{N}$, then $f^{-1}(\cup_i A_i) = \cup_i f^{-1}(A_i)$. So $\cup_i A_i \in \mathcal{F}$
- By the definition of a continuous function, for any open set A, $f^{-1}(A)$ is open, so $f^{-1}(A) \in \mathcal{B}$. Hence $\mathcal{O} \subset \mathcal{F}$
- Therefore, $\mathcal{B} = \sigma(\mathcal{O}) \subset \mathcal{F}$. This implies that f is Borel measurable

Page 20 in Lecture 5: Properties of natural exponential families

Let $\mathcal{P}$ be a natural exponential family with p.d.f.

$$f_\eta(x) = \exp\{\eta^\top T(x) - \zeta(\eta)\} h(x). \tag{3}$$

Let $T = (Y, U)$ and $\eta = (\vartheta, \varphi)$, where Y and ϑ have the same dimension. Then Y has the p.d.f.

$$f_\eta(y) = \exp\{\vartheta^\top y - \zeta(\eta)\} \tag{4}$$

w.r.t. a σ-finite measure λ_φ depending on φ.

Proof

- Let $\eta_0 = (\vartheta_0, \varphi_0)$ be a point of the natural parameter space.
- Then, by chain rule of R-N derivative,

$$\begin{aligned}\frac{\mathrm{d}P_\eta}{\mathrm{d}P_{\eta_0}}(x) &= \exp\{\zeta(\eta_0) - \zeta(\eta)\}\exp\{(\eta - \eta_0)^\top T(x)\} \\ &= \exp\{\zeta(\eta_0) - \zeta(\eta)\}\exp\{(\vartheta - \vartheta_0)^\top Y(x) + (\varphi - \varphi_0)^\top U(x)\}\end{aligned}$$

- For any $B \in \mathcal{B}^d$ where d is the dimension of Y, we have

$$\begin{aligned}&P_\eta(Y \in B) \\ &= E_\eta\left(I_{Y \in B}\right) \\ &= E_{\eta_0}\left[I_{Y \in B} \times \frac{\mathrm{d}P_\eta}{\mathrm{d}P_{\eta_0}}\right] \\ &= E_{\eta_0}\left[I_{Y \in B} \times \exp\{\zeta(\eta_0) - \zeta(\eta)\}\exp\{(\vartheta - \vartheta_0)^\top Y + (\varphi - \varphi_0)^\top U\}\right]\end{aligned}$$

Proof (Cont.)

By tower property of conditional expectation, we have

$$\begin{aligned} & P_\eta(Y \in B) \\ = & E_{\eta_0}\Big[I_{Y \in B} \times \exp\{\zeta(\eta_0) - \zeta(\eta)\} \times \exp\{(\vartheta - \vartheta_0)^\top Y\} \\ & \quad \times E_{\eta_0}[\exp\{(\varphi - \varphi_0)^\top U\} \mid Y] \Big] \end{aligned}$$

- Let $\mu_\eta(B) = P_\eta(Y \in B)$ for any Borel set B in the range of Y
- Define a measure λ_φ on the space of Y by $\frac{\mathrm{d}\lambda_\varphi}{\mathrm{d}\mu_{\eta_0}}(y) = \exp\{\zeta(\eta_0) - \vartheta_0^\top y\} \times E_{\eta_0}[\exp\{(\varphi - \varphi_0)^\top U\} \mid Y = y]$

We conclude that

$$\mu_\eta(B) = \int I_{y \in B} \times \exp\{-\zeta(\eta)\} \times \exp\left(\vartheta^\top y\right) \, \mathrm{d}\lambda_\varphi(y) \tag{5}$$

Tutorial

1. Let $X_1, \ldots, X_n$ be i.i.d. from a uniform distribution on $(-\theta, \theta)$, where $\theta > 0$ is an unknown parameter.
   (a). Find a minimal sufficient statistic T.
   (b). Define
   $$V = \frac{\bar{X}}{\max_i X_i - \min_i X_i}$$
   where $\bar{X}$ is the sample mean. Are T and V independent?

2. An object with weight θ is weighed on scales with different precision. The data $X_1, \ldots, X_n$ are independent, with $X_i \sim N\left(\theta, \sigma^2\right), i = 1, \ldots, n$ with the standard deviation σ known. Consider the absolute deviation loss $L(\theta, a) = |\theta - a|$.
   (a). What is the risk of the naive estimator X_1?
   (b). Use Rao-Blackwell theorem to find a better estimator.

3. Consider an estimation problem with a parametric family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ and the squared error loss. If $\theta_0 \in \Theta$ satisfies that $P_\theta \ll P_{\theta_0}$ for any $\theta \in \Theta$, show that the estimator $T \equiv \theta_0$ is admissible.

Exercise 1

Let $X_1, \ldots, X_n$ be i.i.d. from a uniform distribution on $(-\theta, \theta)$, where $\theta > 0$ is an unknown parameter.

(a). Find a minimal sufficient statistic T.

(b). Define

$$V = \frac{\bar{X}}{\max_i X_i - \min_i X_i}$$

where $\bar{X}$ is the sample mean. V and T are NOT independent.

Solution: Part (a)

- The joint p.d.f. is $\prod_i [\theta^{-1} I_{(-\theta,\theta)}(x_i)] = \theta^{-n} I_{(x_{(n)},\infty)}(\theta) I_{(-x_{(1)},\infty)}(\theta)$
- Let $T(x) = (x_{(1)}, x_{(n)})$. It is sufficient
- If x and y are two sample points such that

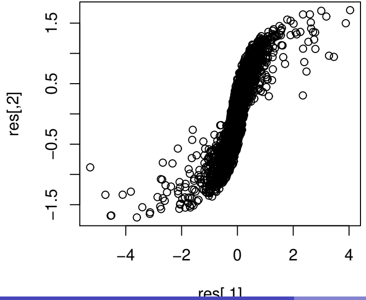
$$f_\theta(x) = f_\theta(y)\phi(x,y), \forall \theta > 0, \tag{6}$$

  where ϕ is a bivariate function, then we can show $T(x) = T(y)$ (see page 21 in Lecture 7). So T is minimal sufficient by Theorem C

Part (b) of the original problem, which asks to show V and T are independent, is wrong.

One can disprove the original problem by doing the following simulation:

- Repeatedly do the following for 5000 times
  - Sample X_i from Unif(-1,1) with $n = 4$
  - Compute $T(X)$ and $V(X)$
  - Record the values of $R_1 = V$ and $R_2 = X_{(n)} + X_{(1)}$.
- Look at the scatter plot of (R_1, R_2), and compute their correlation, which is as high as 0.8



Exercise 2

An object with weight θ is weighed on scales with different precision. The data $X_1, \ldots, X_n$ are independent, with $X_i \sim N\left(\theta, \sigma^2\right), i = 1, \ldots, n$ with the standard deviation σ known. Consider the absolute deviation loss $L(\theta, a) = |\theta - a|$.
(a). What is the risk of the naive estimator X_1?
(b). Use Rao-Blackwell theorem to find a better estimator.

Proof: Part (a): By direct calculation, the risk is
$R_{X_1}(\theta) = \sigma E|Z| = \sigma\sqrt{2/\pi}$, where $Z \sim N(0,1)$
Part (b):

- We know that $T = \sum X_i$ is a sufficient statistic
- By symmetry, we can show $E(X_1 \mid T) = E(X_i \mid T)$ (see Tutorial Q1 in Lecture 5)
- Therefore, $E(X_1 \mid T) = \frac{1}{n}T$
- By Rao-Blackwell theorem and the strict convexity of L, $\frac{1}{n}T$ is a better estimator
- Note that $\frac{1}{n}T \sim N(\theta, \sigma^2/n)$, we have

$$R_{\frac{1}{n}T}(\theta) = \sigma\sqrt{2/(n\pi)} = R_{X_1}(\theta)/\sqrt{n}$$

Exercise 3

Consider an estimation problem with a parametric family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ and the squared error loss. If $\theta_0 \in \Theta$ satisfies that $\mathcal{P}$ is dominated by P_{θ_0}, show that the estimator $T \equiv \theta_0$ is admissible

Proof: We need to show : there is no other rule better than T

- Let U be an estimator of θ such that $R_U(\theta) = E_\theta(U - \theta)^2 \leq R_T(\theta)$ for all θ
- Then $R_U\left(\theta_0\right) \leq R_T(\theta_0) = 0$
- So $E_{\theta_0}\left(U - \theta_0\right)^2 = 0$
- Therefore, $U = \theta_0$, P_{θ_0}-a.s.
- Since $P_\theta \ll P_{\theta_0}$ for any $\theta \in \Theta$, we conclude that $U = \theta_0 = T$ a.s. $\mathcal{P}$