Lecture 10: Loss Function Optimality

Mathematical Statistics II, MATH 60062/70062

Tuesday February 22, 2022

Reference: Casella & Berger, 7.3.4

Loss function optimality

Our evaluations of point estimators have been based on their mean squared error (MSE) performance.

MSE is a special case of a **loss function**.

The study of performance, and the optimality of estimators evaluated through loss functions is called **decision theory**.

Decision theory

After the data $\boldsymbol{X} = \boldsymbol{x}$ are observed, where $\boldsymbol{X} \sim f_{\boldsymbol{X}}(\boldsymbol{x} \mid \theta)$, $\theta \in \Theta$, a *decision* regarding θ is made.

- In point estimation, the decision is a point estimate of θ.
- In hypothesis testing, the decision is an assertion about θ.
- In interval estimation, the decision is an interval estimate.

The set of allowable decisions is the **action space**.

- In estimation, the action space should be some subset of the parameter space.
- In testing, the action space is defined by the null/alternative hypotheses.

Loss functions in point estimation

If an action (i.e., estimate) a is close to θ, then the decision a is reasonable, and little loss is incurred. If a is far from θ, then a large loss is incurred.

The loss function is a nonnegative function that increases as the distance between a and θ increases. If θ is real-valued, two common loss functions are

- Absolute error loss: $L(\theta, a) = |a - \theta|$
- Squared error loss: $L(\theta, a) = (a - \theta)^2$

Squared error loss gives relatively more penalty for large discrepancies, and absolute error loss gives relatively more penalty for small discrepancies.

Considering the consequences of various errors in estimation for different values of θ, other loss functions may be used:

- One penalizes overestimation more than underestimation

$$L(\theta, a) = \begin{cases} (a-\theta)^2 & \text{if } a < \theta \\ 10(a-\theta)^2 & \text{if } a \geq \theta \end{cases}.$$

- One penalizes errors in estimation more if θ is near zero than if $|\theta|$ is large

$$L(\theta, a) = \frac{(a-\theta)^2}{|\theta|+1}.$$

Risk functions

In decision theoretic analysis, the quality of an estimator is quantified through its **risk function**.

For an estimator $\delta(\boldsymbol{x})$ of θ, the risk function is

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

Since the true value of θ is unknown, the estimator that has a small value of $R(\theta, \delta)$ for all values of $\theta \in \Theta$ is desirable.

When comparing two estimators δ_1 and δ_2, if $R(\theta, \delta_1) < R(\theta, \delta_2)$ for all $\theta \in \Theta$, then δ_1 is preferred.

In previous discussions, we have considered mean squared error

$$R(\theta, \delta) = \mathrm{Var}_\theta[\delta(\boldsymbol{X})] + \mathrm{Bias}_\theta^2[\delta(\boldsymbol{X})]$$

Let $\mathcal{D}$ represent the set of allowable decisions (estimators). In decision theoretic analysis, it would be atypical to restrict $\mathcal{D}$ to be the set of only unbiased estimators.

A decision theoretic analysis would be more comprehensive in that both the variance and the bias are considered simultaneously.

Bayesian approach to loss function optimality

In a Bayesian analysis, we would use a prior distribution to compute an average risk known as the **Bayes risk**.

For $\boldsymbol{X} \sim f_{\boldsymbol{X}}(\boldsymbol{x} \mid \theta)$ and $\theta \sim \pi$, the Bayes risk of a decision rule is

$$\begin{aligned}\int_{\Theta} R(\theta, \delta)\pi(\theta)d\theta &= \int_{\Theta} \left(\int_{\mathcal{X}} L(\theta, \delta(\boldsymbol{x})) f_{\boldsymbol{X}}(\boldsymbol{x} \mid \theta) d\boldsymbol{x} \right) \pi(\theta) d\theta \\ &= \int_{\mathcal{X}} \left(\int_{\Theta} L(\theta, \delta(\boldsymbol{X})) f(\theta \mid \boldsymbol{x}) d\theta \right) m(\boldsymbol{x}) d\boldsymbol{x},\end{aligned}$$

where $f(\theta \mid \boldsymbol{x})$ is the posterior distribution of θ and $m(\boldsymbol{x})$ is the marginal distribution of $\boldsymbol{X}$.

An estimator that yields the smallest value of the Bayes risk is called the **Bayes rule with respect to a prior** π, denoted by δ^{π}.

Expected posterior loss

In the Bayes risk

$$\begin{aligned}\int_{\Theta} R(\theta,\delta)\pi(\theta)d\theta &= \int_{\Theta}\left(\int_{\mathcal{X}} L(\theta,\delta(\boldsymbol{x}))f_{\boldsymbol{X}}(\boldsymbol{x}\mid\theta)d\boldsymbol{x}\right)\pi(\theta)d\theta \\ &= \int_{\mathcal{X}}\left(\int_{\Theta} L(\theta,\delta(\boldsymbol{X}))f(\theta\mid\boldsymbol{x})d\theta\right)m(\boldsymbol{x})d\boldsymbol{x},\end{aligned}$$

the expected loss with respect to the posterior distribution is called the **expected posterior loss**.

The expected posterior loss is a function of $\boldsymbol{x}$, not of θ.

For each $\boldsymbol{x}$, if we choose the action $\delta(\boldsymbol{x})$ to minimize the posterior expected loss, we will minimize the Bayes risk.

Two Bayes rules

Consider a point estimation problem for $\theta \in \Theta$.

1. For squared error loss, the posterior expected loss is

$$\int_{\Theta} (a-\theta)^2 f(\theta \mid \boldsymbol{x}) d\theta = E[(\theta - a)^2 \mid \boldsymbol{X} = \boldsymbol{x}]$$

   where the variation in θ is modeled by the posterior distribution $f(\theta \mid \boldsymbol{x})$. This expected value is minimized by

$$\delta^{\pi}(\boldsymbol{x}) = E[\theta \mid \boldsymbol{x}].$$

   See example 2.2.6 of the book

   So the Bayes rule is the mean of the posterior distribution.

2. For absolute error loss, the posterior expected loss is

$$E(|\theta - a| \mid \boldsymbol{X} = \boldsymbol{x}),$$

   which is minimized by choosing the median of the posterior.

Binomial Bayes estimates

Suppose that $X_1, \ldots, X_n$ are iid $\mathrm{Bern}(\theta)$, where the prior distribution on θ is $\mathrm{Beta}(a, b)$. Let $T = \sum_{i=1}^n X_i$. The posterior distribution of θ depends on the sample through the observed value $T = t$ and is $\mathrm{Beta}(t + a, n - t + b)$.

- The Bayes estimator of θ for squared error loss is

$$\delta^\pi(t) = E(\theta \mid t) = \frac{t + a}{a + b + n}.$$

- For absolute error loss, the Bayes estimator be the number η satisfying

$$\int_0^\eta \frac{\Gamma(a + b + n)}{\Gamma(t + a)\Gamma(n - t + b)} \theta^{t+a-1}(1 - \theta)^{n-t+b-1} d\theta = \frac{1}{2}.$$

  That is, η is the median of the posterior distribution, which can be found using numerical methods.