Fundamental Theory of Statistical Inference

G. Alastair Young

Department of Mathematics
Imperial College London

LTCC, 2017

Formulation

Elements of a formal decision problem:

(1) Parameter space Ω_θ. Represents the set of possible unknown states of nature.

(2) Sample space $\mathcal{Y}$. Typically have n observations, so a generic element of the sample space is $y = (y_1, ..., y_n) \in \mathbb{R}^n$.

(3) Family of distributions. $\{\mathbb{P}_\theta(y),\ y \in \mathcal{Y},\ \theta \in \Omega_\theta\}$. Generally consists of a family $f(y;\theta)$ of probability mass functions or density functions of Y.

(4) Action space $\mathcal{A}$. Set of all actions or decisions available.

Example 1. Hypothesis testing problem, two hypotheses H_0 and H_1, $\mathcal{A} = \{a_0, a_1\}$, a_0 represents accepting H_0, a_1 represents accepting H_1.

Example 2. In point estimation typically have $\mathcal{A} \equiv \Omega_\theta$.

(5) Loss function L. Function $L : \Omega_\theta \times \mathcal{A} \to \mathbb{R}$ links the action to the unknown parameter: if we take action $a \in \mathcal{A}$ when the true state of nature is $\theta \in \Omega_\theta$, then we incur a loss $L(\theta, a)$.

(6) Decision rule d. Function $d : \mathcal{Y} \rightarrow \mathcal{A}$. Each point $y \in \mathcal{Y}$ is associated with a specific action $d(y) \in \mathcal{A}$.

Example 1. If $\bar{y} \leq 5.7$, accept H_0, otherwise accept H_1. So, $d(y) = a_0$ if $\bar{y} \leq 5.7$, $d(y) = a_1$ otherwise.

Example 2. Estimate θ by $d(y) = y_1^3 + 27 sin(\sqrt{y_2})$.

The Risk Function

Risk associated with decision rule d based on random data Y given by

$$R(\theta, d) = \mathbb{E}_\theta L(\theta, d(Y)) = \int_{\mathcal{Y}} L(\theta, d(y)) f(y; \theta) dy.$$

Expectation of loss with respect to distribution on Y, for the particular θ.

Different decision rules compared by comparing their risk functions, as functions of θ. Repeated sampling principle explicitly invoked.

Utility and loss

Utility theory: measure of loss in terms of utility to individual.

If behave rationally, act as if maximising the expected value of a utility function.

Adopt instead various artificial loss functions, such as

$$L(\theta, a) = (\theta - a)^2,$$

the squared error loss function. When estimating a parameter θ, we seek a decision rule $d(y)$ which minimises the mean squared error $\mathbb{E}_\theta\{\theta - d(Y)\}^2$.

Other loss functions

Can consider other loss functions, such as absolute error loss,

$$L(\theta, a) = |\theta - a|.$$

In hypothesis testing, where we have two hypotheses H_0, H_1, and corresponding action space $\mathcal{A} = \{a_0, a_1\}$, the most familiar loss function is

$$L(\theta, a) = \begin{cases} 1 & \text{if } \theta \in H_0 \text{ and } a = a_1 \\ 1 & \text{if } \theta \in H_1 \text{ and } a = a_0 \\ 0 & \text{otherwise.} \end{cases}$$

In this case the risk function is the probability of making a wrong decision:

$$R(\theta, d) = \begin{cases} \mathbb{P}_\theta\{d(Y) = a_1\} & \text{if } \theta \in H_0 \\ \mathbb{P}_\theta\{d(Y) = a_0\} & \text{if } \theta \in H_1. \end{cases}$$

Criteria for a good decision rule

Ideally, find a decision rule d which makes the risk function $R(\theta, d)$ uniformly small for all values of θ.

Rarely possible, so consider a number of criteria which help to narrow down the class of decision rules we consider.

Admissible decision rules

Given two decision rules d and d', we say d strictly dominates d' if $R(\theta, d) \leq R(\theta, d')$ for all values of θ, and $R(\theta, d) < R(\theta, d')$ for at least one θ.

Any decision rule which is strictly dominated by another decision rule is said to be inadmissible. If a decision rule d is not strictly dominated by any other decision rule, then it is admissible.

Admissibility: absence of a negative attribute.

Minimax decision rules

The maximum risk of a decision rule d is defined by

$$\mathrm{MR}(d) = \sup_{\theta} R(\theta, d).$$

A decision rule d is minimax if it minimises the maximum risk:

$$\mathrm{MR}(d) \leq \mathrm{MR}(d') \text{ for all decision rules } d'.$$

So, d must satisfy

$$\sup_{\theta} R(\theta, d) = \inf_{d'} \sup_{\theta} R(\theta, d').$$

In most problems we encounter, the maxima and minima are actually attained.

Minimax principle

The minimax principle says we should use the minimax decision rule.

Protects against worst case, may lead to counterintuitive result.

If minimax rule is not admissible, can find another which is.

Unbiased decision rules

A decision rule d is said to be unbiased if

$$\mathbb{E}_\theta\{L(\theta', d(Y))\} \geq \mathbb{E}_\theta\{L(\theta, d(Y))\} \text{ for all } \theta, \theta'.$$

Suppose the loss function is squared error $L(\theta, d) = (\theta - d)^2$. For d to be an unbiased decision rule, we require $d(Y)$ to be an unbiased estimator in the classical sense.

Discussion

Role of unbiasedness is ambiguous.

As criterion, doesn't depend solely on risk function.

Bayes decision rules

In addition to loss function, specify a prior distribution which represents our prior knowledge of the parameter θ, and is represented by a function $\pi(\theta),\ \theta \in \Omega_\theta$.

If Ω_θ is a continuous parameter space, such as an open subset of $\mathbb{R}^k$ for some $k \geq 1$, usually assume that the prior distribution is absolutely continuous and take $\pi(\theta)$ to be some probability density on Ω_θ. In the case of a discrete parameter space, $\pi(\theta)$ is a probability mass function.

Bayes risk

In the continuous case, the Bayes risk of a decision rule d is defined to be

$$r(\pi, d) = \int_{\theta \in \Omega_\theta} R(\theta, d)\pi(\theta) d\theta.$$

In the discrete case, integral is replaced by a summation.

Bayes rule

A decision rule d is said to be the Bayes rule (with respect to a given prior $\pi(\cdot)$) if it minimises the Bayes risk: if

$$r(\pi, d) = \inf_{d'} r(\pi, d') = m_\pi \text{ say.}$$

The Bayes principle says we should use the Bayes decision rule.

Some other definitions

Sometimes the Bayes rule is not defined because the infimum is not attained for any decision rule d. However, in such cases, for any $\epsilon > 0$ we can find a decision rule d_ϵ for which

$$r(\pi, d_\epsilon) < m_\pi + \epsilon$$

and in this case d_ϵ is said to be ϵ-Bayes (with respect to the prior distribution $\pi(\cdot)$).

A decision rule d is said to be extended Bayes if, for every $\epsilon > 0$, we have that d is ϵ-Bayes with respect to some prior (which need not be the same for different ϵ).

Randomised Decision Rules

Suppose we have I decision rules $d_1, ..., d_I$ and associated probability weights $p_1, ..., p_I$ ($p_i \geq 0$ for $1 \leq i \leq I$, $\sum_i p_i = 1$).

Define $d^* = \sum_i p_i d_i$ to be the decision rule "select d_i with probability p_i": imagine using some randomisation mechanism to select among the decision rules $d_1, ..., d_I$ with probabilities $p_1, ..., p_I$, and then, having decided in favour of some d_i, carry out the action $d_i(y)$ when y is observed.

d^* is a randomised decision rule.

Risk of randomised rule

For a randomised decision rule d^*, the risk function is defined by averaging across possible risks associated with the component decision rules:

$$R(\theta, d^*) = \sum_{i=1}^{I} p_i R(\theta, d_i).$$

Randomised decision rules may appear to be artificial, but minimax solutions may well be of this form.

Finite Decision Problems

Suppose parameter space is a finite set, $\Omega_\theta = \{\theta_1, ..., \theta_t\}$ for some finite t, with $\theta_1, \ldots, \theta_t$ specified..

Notions of admissible, minimax and Bayes decision rules can be given a geometric interpretation.

Define the risk set to be a subset S of $\mathbb{R}^t$, generic point consists of the t-vector $(R(\theta_1, d), ..., R(\theta_t, d))$ associated with a decision rule d.

Assume the space of decision rules includes all randomised rules.

The risk set S is a convex set. Minimax rules etc. can often be identified by drawing S as subset of $\mathbb{R}^t$

Finding minimax rules in general

Theorem 2.1 If δ_n is Bayes with respect to prior $\pi_n(\cdot)$, and $r(\pi_n, \delta_n) \to C$ as $n \to \infty$, and if $R(\theta, \delta_0) \leq C$ for all $\theta \in \Omega_\theta$, then δ_0 is minimax.

[This includes the case where $\delta_n = \delta_0$ for all n and the Bayes risk of δ_0 is exactly C.]

A decision rule d is an equaliser decision rule is $R(\theta, d)$ is the same for every value of θ.

Theorem 2.2 An equaliser decision rule δ_0 which is extended Bayes must be minimax.

Admissibility of Bayes rules

Bayes rules are nearly always admissible.

Theorem 2.3 Assume that Ω_θ is discrete, $\Omega_\theta = \{\theta_1, \ldots, \theta_t\}$ and that the prior π gives positive probability to each θ_i. A Bayes rule with respect to π is admissible.

Theorem 2.4 If a Bayes rule is unique, it is admissible.

Theorem 2.5 Let Ω_θ be a subset of the real line. Assume that the risk functions $R(\theta, d)$ are continuous in θ for all decision rules d. Suppose that for any $\epsilon > 0$ and any θ the interval $(\theta - \epsilon, \theta + \epsilon)$ has positive probability under the prior π. Then a Bayes rule with respect to π is admissible.