STAT 610: Statistical Inference

Lecture 1. Overview

Instructor:
Harry Zhou: huibin.zhou@yale.edu. Office hours: Thursday 5:00-6:30pm or by appointment in Room 204, 24 Hillhouse Ave, Department of Statistics.

References:
We will use *Theory of Point Estimation* as the main textbook. I will give weekly lecture notes, which may cover topics beyond the textbook, such as basic lower bound arguments, false discovery rate, frequentist justifications of Bayesian methods, etc.

- Erich L. Lehmann, George Casella, *Theory of Point Estimation*, 2nd Edition, Springer (1998).
- George Casella, Roger L. Berger, *Statistical Inference* Edition: text binding.

Location and time:
Room 107, 24 Hillhouse Ave, 10:30-11:45AM, Tuesday and Thursday.

Teaching assistant:
Dana Yang: xiaoqian.yang@yale.edu. Office hours: Friday 9:00-11:00AM.

Grade:
Weekly Homework: 30%. **Due every Friday before noon.**
Take-home Midterm: 30%.
Take-home Final Exam: 40%.

Basic concepts:
Wald (1939), Contributions to the Theory of Statistical Estimation and Testing Hypotheses, *Ann. Math. Stat.*. This paper introduced much of the landscape of modern decision theory, including loss functions, risk functions, admissible decision rules, prior, Bayes decision rules, and minimax decision rules. Wolfowitz described this paper as: " ... probably Wald's most important single paper". The phase "decision theory" was possibly first used by Lehmann.

Example 1: Observe a normally distributed p-dimensional random vector X,

$$X \sim N\left(\theta, I_{p\times p}\right),$$

where θ is unknown parameter. We assume that the parameter space is $\Theta \subset \mathbb{R}^p$. Our goal is to estimate the mean vector θ, or more generally $g(\theta)$, a functional of θ. Let $\delta\left(X\right)$ be an estimator of θ.

Estimator. An estimator is a real-valued function of the observed values or data (in sample space $\mathcal{X}$). It is used to estimate an *estimand* $g(\theta)$, a real-valued function of the parameter (in parameter space Ω). In Example 1, $g(\theta) = \theta$, or $g(\theta) = \|\theta\|^2 = \sum_{i=1}^{p} \theta_i^2$. For now on in this lecture let's assume $g(\theta) = \theta$.

Loss function. We assume $L(\theta, \delta(X)) \geq 0$, and $L(\theta, \delta) = 0$ implies $\delta(X) = \theta$. A commonly used loss function is

$$L(\theta, \delta) = \|\theta - \delta(X)\|^2 = \sum_{i=1}^{p} (\delta_i - \theta_i)^2 .$$

Risk function. The risk function is used to measure how well the estimator does on average

$$R(\theta, \delta) = E_{X|\theta} L(\theta, \delta) .$$

Question: Is there a *uniformly optimal* estimator over all $\theta \in \Theta$? The answer is usually no. Can we find $T(x)$ such that

$$E_{X|\theta} \|\theta - T(X)\|^2 \leq E_{X|\theta} \|\theta - \delta(X)\|^2$$

for all $\delta(X)$ and all θ?

Unbiased estimation. In Chapter 2, we consider unbiased estimators. An estimator $\delta(X)$ is called *unbiased* if $E_{X|\theta}\delta(X) = \theta, \forall\theta$. It is possible to find an estimator with the smallest risk among all unbiased estimators.

Equivariant estimation. In Chapter 3, we discuss the best equivariant estimator. The topic is less important in modern statistics. If

$$\delta(X + a) = a + \delta(X),$$

we call δ is *location equivariant*.

Bayesian estimation. In Chapter 4, we consider the *average risk*. Assume θ has a prior distribution Π. If the corresponding density is $\pi(\theta)$, $\int E_{X|\theta} \|\delta(X) - \theta\|^2 \pi(\theta) d\theta$ is an average risk. A question we would like to ask is: How we find an estimator $T(X)$ s.t.

$$\int E_{X|\theta} \|T(X) - \theta\|^2 \pi(\theta) d\theta \leq \int E_{X|\theta} \|\delta(X) - \theta\|^2 \pi(\theta) d\theta$$

for all $\delta(X)$?

Admissible. An estimator δ is called to be inadmissible if there is an estimator δ' such that

$$R(\theta, \delta') \leq R(\theta, \delta) \text{ for all } \theta \in \Theta, \text{ and } R(\theta, \delta') < R(\theta, \delta) \text{ for some } \theta \in \Theta.$$

An estimator is admissible if it is not inadmissible. Wald (1939) operated with Bayes solution. A brave man?

Remark: In 1939, Wald "proved" the admissibility of the estimator X. Stein received his Ph.D. in 1947 from Columbia under Wald on sequential analysis.

Inspired by Savage, Stein Started to realize that the inadmissibility was perhaps true.

Stein (1956), Inadmissibility of the Usual Estimator for the Mean of a Multivariate Norma Distribution. Proc. 3^{rd} Berk Symp Math. Stat. Prob.

James and Stein (1961), Estimation with quadratic loss. Proc. 4^{th} Berkeley Symp. Math. Statist. Prob., 1.

James-Stein estimator

$$\delta_{J-S}(X) = \left(1 - \frac{c}{\|X\|^2}\right) X,\ c > 0.$$

This estimator is minimax if and only if $0 \le c \le 2(p-2)$. Later this semester we will show that for $0 < c < 2(p-2)$,

$$E_{X|\theta} \|\delta_{J-S}(X) - \theta\|^2 < E_{X|\theta} \|X - \theta\|^2,$$

which implies X is not admissible.

Minimaxity. In Chapter 5, we consider *minimizing the worse risk.* An estimator δ^* is minimax if

$$R(\Theta, \delta^*) = \inf_{\delta} \sup_{\theta \in \Theta} R(\theta, \delta).$$

Why minimax? In Wald's book he states on page 27: "Nevertheless, since Nature's choice is unknown to the experimenter, it is perhaps not unreasonable to for experimenter to behave as if Nature wanted to maximize the risk. But, even if one is not willing to take this attitude, the theory of games remains of fundamental importance ...".

There may be few statisticians who actively supports the minimax principle as a prescription for action. However the minimax idea has been an essential foundation for advances in many areas of statistical research: asymptotic theory and methodology, hierarchical models, robust estimation, optimal design, and nonparametric function estimation.

Hypothesis Testing: Testing is closely connected to estimation. In modern statistics, we explore this relationship a lot.

Asymptotics. In Chapter 6, we consider a setting that the sample size grows. We will study asymptotic properties of maximum likelihood estimation and Bayesian estimation.