STA732

Statistical Inference

Lecture 01: Course Introduction

Yuansi Chen

Spring 2023

Duke University

https://www2.stat.duke.edu/courses/Spring23/sta732.01/

Goal of Lecture 01

- Logistics
- Introduce “the problem”
- Discuss what it means to have “the best” estimator?
- (If time permits) Review of measure theory basics

Logistics

Coordinates

- **Instructor:** Yuansi Chen yuansi.chen@duke.edu
- **TA:** Christine Shen yueming.shen@duke.edu
- **Course websites:**
  - Main: https://www2.stat.duke.edu/courses/Spring23/sta732.01/
  - Sakai

Lectures and Office hours

- **Lectures**: Monday and Wednesday 3:30-4:45pm in Old Chem 025
- **Office hours**:
  - Yuansi: MW 4:45-5:30
  - Christine: see website
- Ed Discussion on Sakai

Sakai tools

- Announcements
- Zoom Meetings (for possible online office hours)
- Resources (for HW problem sets)
- Ed Discussion (for online discussions)
- Gradescope (for HW submission and exam grading)

Textbooks

- Keener, *Theoretical Statistics: Topics for a Core Course*, 2010
- Lehmann and Casella, *Theory of Point Estimation*, 1998
- Lehmann and Romano, *Testing Statistical Hypotheses*, 2005

All available online via Duke library website

Grading

- Weekly homework (due on Wednesdays at 11am)
- One midterm + one final

Homework	25%
Midterm	25%
Final	45%
Participation	5%

Scribing

Everyone is required to scribe at least one lecture note. Please sign up via the link in Sakai.

Policies

Check website

- Duke Community Standard
  - I will not lie, cheat, or steal in my academic endeavors
  - I will conduct myself honorably in all of my endeavors
  - I will act if the standard is compromised
- Plagiarism
  - Can use online resources, but make sure you understand in your own language and make sure to cite them (code or theory)
  - Answer sharing between groups or individuals are not allowed
- Homework Policy:
  - Late HW: No homework more than two full days (48 hours) late will be accepted. Each late day will result in a one-level down-grade (10% off) of that HW
  - Regrade requests on Gradescope within 2 days
  - Drop the HW with the lowest score for final grade
- Exam Policy: no makeup exams

HW0 released

- Designed for trying out Gradescope on Sakai
- due Wednesday 18 at 11am
- will not be counted into final grade

The statistical inference problem

Statistical inference in a dictionary

Oxford Dictionary of Statistics or Wikipedia:

“Statistical inference is the process of using data analysis to infer properties of an underlying distribution of probability”

Statistical experiment

Statistical experiment

A statistical experiment is a procedure/process that generates a collection of data, **X**

- For example, a coin tossing experiment: tossing a coin n times.

Statistical experiment

Statistical experiment

A statistical experiment is a procedure/process that generates a collection of data, **X**

- For example, a coin tossing experiment: tossing a coin n times.

Sample space

The set of possible data values is called the sample space **S**

- For example, in the coin tossing experiment, $\mathbf{S} = \{0, 1\}^n$, the sample space contains all length n string with 0s and 1s

Statistical model

Statistical model

A statistical model is a family of possible distributions $\{P_\theta, \theta \in \Omega\}$ for $\mathbf{X}$, where Ω is called the parameter space

- Note that the family can be very small (e.g. a single distribution) or very large (e.g. all absolutely continuous distributions)
- Bayesian also puts assumption on θ (we will deal later)
- A model is in essence a collection of assumptions regarding the sampling distribution of the data

Event in the sample space

- $E \subset \mathbf{S}$ is called an event
- Each distribution in a model can specify the probability of an event

$$P_\theta(E) = \text{Prob}_\theta(\mathbf{X} \in E)$$

Example: coin tossing experiment

- **Data: X** $= (X_1, X_2, \ldots, X_n)$ where

$$X_i = \begin{cases} 1 & \text{if the } i\text{th toss is H} \\ 0 & \text{if the } i\text{th toss is T} \end{cases}$$

- **Statistical model:** contains all joint distributions of n independent Bernoulli distribution with equal head probability $\theta, \theta \in [0, 1]$

$$P_\theta(X_1 = x_1, \ldots, X_n = x_n) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$$

Statistical inference

Inference

Inference about $g(\theta)$ (estimand) is making an "educated guess" about $g(\theta)$ based on the data

- for example, blindly guessing θ to be 0 is a type of inference
- guessing θ to be a real number between 0 and 1 is also an example of inference

Common types of statistical inference problems

1. Point estimation
2. Hypothesis testing
3. Interval estimation (confidence intervals or credible regions)
4. Prediction

Common types of statistical inference problems (2)

1. **Point estimation:** an estimator is a statistic (a function of the data) for the purpose of guessing the value for some $g(\theta)$, which is hopefully close to $g(\theta)$

Common types of statistical inference problems (2)

1. **Point estimation:** an estimator is a statistic (a function of the data) for the purpose of guessing the value for some $g(\theta)$, which is hopefully close to $g(\theta)$
2. **Hypothesis testing:** let $\Omega = \Omega_0 \cup \Omega_1$ be a disjoint union. Ask whether

$$H_0 : \theta \in \Omega_0 \text{ or } H_1 : \theta \in \Omega_1$$

Common types of statistical inference problems (3)

3. **Interval estimation:** suppose $g(\theta) \in \mathbb{R}$. We want an interval that contains $g(\theta)$ "with high probability"
   - A level $(1-\alpha) * 100\%$ confidence interval $[l(\mathbf{X}), u(\mathbf{X})]$:

$$P_\theta \left(g(\theta) \in [l(\mathbf{X}), u(\mathbf{X})] \right) \geq 1 - \alpha$$

$(1-\alpha)$ is called the significance level or coverage level

Common types of statistical inference problems (3)

3. **Interval estimation:** suppose $g(\theta) \in \mathbb{R}$. We want an interval that contains $g(\theta)$ "with high probability"
   - A level $(1-\alpha) * 100\%$ confidence interval $[l(\mathbf{X}), u(\mathbf{X})]$:

$$P_\theta \left(g(\theta) \in [l(\mathbf{X}), u(\mathbf{X})] \right) \geq 1 - \alpha$$

   $(1-\alpha)$ is called the significance level or coverage level

4. **Prediction:** what would a new data point look like?

Based on our definition, the requirement of doing inference is quite low. We need a notion of “good inference” to compare inference methods, and to rule out the clearly useless inference methods!

What does it mean to have “the best” estimator?

Take point estimation as an example

Objective of Point estimation:

Construct a statistics $T(\mathbf{X})$ that is "close" to $g(\theta)$.

- What is a formal notion of "closeness"?
- Introduce a loss function

$$L(\theta, d) = \text{the loss incurred when estimating } g(\theta) \text{ by } d$$

Note that d is taken to our estimator which is a statistic (depends on $\mathbf{X}$)

Examples of loss functions

- Squared error loss

$$L(\theta, d) = (d - g(\theta))^2$$

- L_p loss

$$L(\theta, d) = |d - g(\theta)|^p$$

- ϵ-step error loss, $\epsilon > 0$

$$L(\theta, d) = \begin{cases} 1 & \text{if } |d - g(\theta)| > \epsilon \\ 0 & \text{otherwise} \end{cases}$$

- In general, the loss does not have to be symmetric

$$L(\theta, d) = \begin{cases} 3(d - g(\theta)) & \text{if } d - g(\theta) > 0 \\ -(d - g(\theta)) & \text{otherwise} \end{cases}$$

We assume for simplicity the minimum value is taken at $g(\theta)$

Example: a normal experiment

- **Data:** $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ i.i.d. $\mathcal{N}(\theta, 1)$
- **Estimand:** $g(\theta) = \theta$
- **Loss fun:** $L(\theta, d) = (d - \theta)^2$, the squared error loss

Example: a normal experiment

- **Data:** $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ i.i.d. $\mathcal{N}(\theta, 1)$
- **Estimand:** $g(\theta) = \theta$
- **Loss fun:** $L(\theta, d) = (d - \theta)^2$, the squared error loss

If we take $d = \delta(\mathbf{X})$ as our estimator, then the loss $L(\theta, \delta(\mathbf{X}))$

- is random (under repeated experiments)
- depends on the unknown θ

Example: a normal experiment

- **Data:** $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ i.i.d. $\mathcal{N}(\theta, 1)$
- **Estimand:** $g(\theta) = \theta$
- **Loss fun:** $L(\theta, d) = (d - \theta)^2$, the squared error loss

If we take $d = \delta(\mathbf{X})$ as our estimator, then the loss $L(\theta, \delta(\mathbf{X}))$

- is random (under repeated experiments)
- depends on the unknown θ

How do we evaluate the performance of different estimators? say

- sample mean $\bar{X} = \frac{1}{n} \sum X_i$
- sample median $\text{med}(\mathbf{X}) = \text{median}\,(X_1, \ldots, X_n)$

Risk function

The risk function $R(\theta, \delta)$ for an estimator δ is the average loss under repeated experiments (This is the frequentist prespective).

$$R(\theta, \delta) = \mathbb{E}_{\mathbf{X} \sim P_\theta} \left[L(\theta, \delta(\mathbf{X})) \right].$$

In general, we want to find estimators that has "low" risk

but how low is low? low for which θ?

Example: a normal experiment (cont'd)

- The risk of sample mean is

$$R(\theta, \bar{X}) = \mathbb{E}_\theta \left[\left(\bar{X} - \theta \right)^2 \right] = \mathrm{Var}_\theta(\bar{X}) = \frac{1}{n}$$

- The risk of sample median
  - is also constant over all θ
  - is larger than $\frac{1}{n}$ (we will prove later)

In this case, the sample mean is preferred under the squared error loss (and repeated experiments) for all θ. We say the sample mean is uniformly better than the sample median.

A natural question:

Given a loss function, does there exist a uniformly best estimator, which has lower risk than any other estimator over all values of θ?

The answer to the previous question is NO in general!

Proof of NO uniformly best estimator

1. If δ^* exists which is uniformly better, then taking $\delta_c = c$, we must have

$$R(\theta, \delta^*) \leq R(\theta, \delta_c)$$

2. In particular, $R(\theta, \delta^*) \leq R(\theta, g(\theta))$
3. Since $L(\theta, g(\theta))$ is the minimum, $L(\theta, \delta^*(\mathbf{X})) = L(\theta, g(\theta))$ for all θ and all $\mathbf{X}$. This is a degenerate case

In other words, $\delta^*(\mathbf{X}) = \arg\min_\delta L(\theta, \delta)$ no matter what data $\mathbf{X}$ is.

Various approaches to define "good" estimator with "low" risk

Since no uniformly best estimator exist, we must be careful when claiming that "we have the best estimator". We need some compromises or restrictions in defining what is "the best".

Three general approaches

1. Restrict attention to a smaller (but hopefully reasonable) class of estimators (avoid comparison to all estimators)
2. Applying global measures for risk and minimize those, rather than trying to find an estimator with lowest risk at every θ (don't need to be good at every θ)
3. Large sample (asymptotic) approach

I. Restrict attention to a smaller class of estimators

Strategy A: restrict attention to unbiased estimators

- **bias:** $\mathbb{E}_{\mathbf{X}\sim P_\theta}\delta(\mathbf{X}) - g(\theta)$ is the bias of δ
- **UMVU:** Uniformly minimum variance unbiased estimator
- If δ is unbiased, then for square loss

$$\begin{aligned} R(\theta, \delta) &= \mathbb{E}_\theta \left(\delta - g(\theta)\right)^2 \\ &= \mathrm{Var}_\theta(\delta) + \left(\mathbb{E}_\theta \delta - g(\theta)\right)^2 \\ &= \mathrm{Var}_\theta(\delta) \end{aligned}$$

I. Restrict attention to a smaller class of estimators

Strategy A: restrict attention to unbiased estimators

- **bias:** $\mathbb{E}_{\mathbf{X}\sim P_\theta}\delta(\mathbf{X}) - g(\theta)$ is the bias of δ
- **UMVU:** Uniformly minimum variance unbiased estimator
- If δ is unbiased, then for square loss

$$\begin{aligned} R(\theta, \delta) &= \mathbb{E}_\theta \left(\delta - g(\theta)\right)^2 \\ &= \mathrm{Var}_\theta(\delta) + \left(\mathbb{E}_\theta \delta - g(\theta)\right)^2 \\ &= \mathrm{Var}_\theta(\delta) \end{aligned}$$

Strategy B: restrict attention to estimators with certain symmetry

It is sometimes reasonable to require an estimator to be equivariant

$$\delta(X_1 + c, \ldots, X_n + c) = \delta(X_1, \ldots, X_n) + c$$

Strategy A: minimax

- Want to minimize maximum value of risk function, i.e. find δ^* satisfying

$$\sup_{\theta \in \Omega} R(\theta, \delta^*) \leq \sup_{\theta \in \Omega} R(\theta, \delta) \text{ for any other } \delta$$

- Such an estimator is called minimax. It seems rather pessimistic:
  - An estimator might be good for a vast majority of θ, but only bad for a single value of θ. It will not be considered good under the minimax strategy!

Strategy B: minimize the average risk

- Want to minimize the averaged risk under some weight function, i.e. find δ^* to minimize

$$\int R(\theta, \delta) d\Lambda(\theta)$$

  where $\Lambda(\theta)$ is some measure over θ
- If Λ is taken as a probability distribution over the parameter space Ω, then it is called the prior distribution
- Depending on the prior about θ, we weight the risk differently
- Such an estimator is called a Bayes estimator

III. Large sample approach

Intuition for large sample approach: when n tends to infinity, the risk simplifies and we might be able to define which estimator is the best without making too much compromises

Now we can understand why the Lehmann and Casella book (Theory of Point Estimation) is organized as follows

- Preparations
- Unbiasedness
- Equivariance
- Average Risk Optimality
- Minimaxity and Admissibility
- Asymptotic Optimality

What is covered in this course?

- **The first half:** focus on the logic of Lehmann and Casella
- **The second half:** focus on
  - hypothesis testing
  - how the classic studies the maximum likelihood estimator
- But before the first half, we need to have some probability background and to build the basic language
  - Measure theory basics
  - Exponential families
  - Sufficient statistics
  - Rao-Blackwell theorem (a generic way to improve an estimator)

Review of measure theory basics

Measure theory overview

- Measure theory is the foundation of all rigorous statistical theory which is built on top of probability theory
- We will go through the basics, but it is recommended to review Sta 711 textbooks to get a thorough understanding!

Measure

Given a set $\mathcal{X}$, a measure μ maps subsets $A \subseteq \mathcal{X}$ to $[0, \infty)$

- **Example 1:** if $\mathcal{X}$ is countable (e.g. $\mathcal{X} = \mathbf{Z}$), the counting measure $\#(A)$ equals the number of points in A
- **Example 2:** if $\mathcal{X} = \mathbb{R}^n$, the Lebesgue measure is $\lambda(A) = \int \cdots \int_A dx_1 \ldots dx_n = \text{Vol}(A)$

Due to pathological sets, $\lambda(A)$ can only be defined for some subsets $A \subseteq \mathbb{R}^n$. This leads to the introduction of σ-algebra (or σ-field).

σ-algebra

A σ-algebra $\mathcal{F}$ on a set $\mathcal{X}$ is a collection of subsets of $\mathcal{X}$ satisfying

- it includes $\mathcal{X}$ and the empty set
- it is closed under complement
- it is closed under countable unions

- **Example 1:** if $\mathcal{X}$ is countable, $\mathcal{F} = 2^{\mathcal{X}}$ (all subsets)
- **Example 2:** if $\mathcal{X} = \mathbb{R}^n$, $\mathcal{F}$ is the Borel σ-field, $\mathcal{B}$, the smallest σ-algebra that contains all rectangles.

Formal definition of measure under σ-algebra

Given $(\mathcal{X}, \mathcal{F})$ (a measurable space), a measure is any map $\mu : \mathcal{F} \to [0, \infty]$ such that

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i) \text{ for disjoint } A_i \in \mathcal{F}$$

If in addition $\mu(\mathcal{X}) = 1$, then μ is a probability measure

Integrals

We can now define integrals using measures. Intuitively, $\int f d\mu$ means summing f with weights $\mu(A)$ on A

- For counting measure, $\int f(x) d\#(x) = \sum_{x \in \mathcal{X}} f(x)$
- For Lebesgue measure, $\int f(x) d\lambda(x) = \int \cdots \int f(x) dx_1 \ldots dx_n$.
  - Indicator fun $\mathbf{1}_{x \in A}$
  - Simple fun $\sum_i a_i \mathbf{1}_{x \in A_i}$
  - Measurable fun (if pre-image is in $\mathcal{F}$): those can be approximated by simple funs (Theorem 1.8 in Keener)

Densities

Given $(\mathcal{X}, \mathcal{F})$ and two measures μ, P, we say P is absolutely continuous with respect to μ if $P(A) = 0$ whenever $\mu(A) = 0$. Note this as $P \ll \mu$.

- If $P \ll \mu$, we can define the density function

$$p = \frac{dP}{d\mu},$$

  where $P(A) = \int_A p(x) d\mu(x)$. This is also called Radon-Nikodym derivative
- If μ is the counting measure, then p is a probability mass function. If μ is the Lesbegue measure, then p is a probability density function

According to the definition, the density function is not unique but agrees almost everywhere

Probability space

A probability space is the triple $(\Omega, \mathcal{F}, P)$

- Sample space Ω, which is the set of all possible outcomes
- Event space $\mathcal{F}$, $A \subset \mathcal{F}$ is called an **event**
- Probability function P, $P(A)$ is the probability of A

Random variable

A random variable is a function $Y : \Omega \to \mathcal{X}$

- We say Y has distribution Q (or $Y \sim Q$) if

$$P(Y \in B) = P(\{w : Y(w) \in B\}) = Q(B)$$

for $B \in \mathcal{F}$

Expectation

The expectation is an integral with respect to P

$$\mathbb{E}[Y] = \int_{\Omega} Y(w) dP(w) = \int x dQ(x).$$

Need to know more about measure theory?

- More in Keener Chap. 1
- More in Sta 711

Summary

What we have covered

- Statistical inference problem
- Intuitively how to argue for the best estimator

Summary

What we have covered

- Statistical inference problem
- Intuitively how to argue for the best estimator

What is next lecture?

- Exponential families

Thank you