Modern Statistical Inference Methods

Spring 2026

Instructor: Xintao Xia (xintaox@zju.edu.cn)

Lecture: Tuesdays, 10:00–12:25

1 Course Overview

This course develops a unified inference toolkit for real-world data where the data-collection mechanism is nontrivial: **survey sampling**, **missing data**, and **causal inference**. A recurring theme is that valid inference requires explicit reasoning about the *mechanism*—sampling design, nonresponse/missingness, or treatment assignment—and that modern methods (IPW, doubly robust estimators, semiparametric efficiency, ML-assisted inference) can be understood within a common framework. To motivate these ideas, we take a *historical perspective*, tracing how *classical problems and early debates* in statistics led to today's emphasis on *valid inference, efficiency, and robustness*, and using these milestones to cultivate *statistical thinking*.

2 Prerequisites

- Probability and mathematical statistics (LLN/CLT, basic likelihood theory)
- Linear regression and matrix algebra
- Asymptotic statistics (optional, but recommended)
- An open mind and a willingness to think critically, participate actively, and share ideas

3 Learning Objectives

By the end of the course, students will be able to:

1. Conduct design-based inference under common survey designs (SRS, stratified, cluster, PPS).
2. Construct variance estimators for complex surveys (linearization and replicate-weight methods).
3. Diagnose and handle missing data under MCAR/MAR/MNAR; implement EM and multiple imputation.
4. Estimate causal effects under potential outcomes and (optionally) graphical perspectives.
5. Implement propensity-score methods, IPW, AIPW/DR.
6. Use ML for nuisance estimation while maintaining valid inference (sample-splitting / cross-fitting).

4 Recommended References

Survey sampling

- Carl-Erik Särndal, Bengt Swensson, Jan Wretman, *Model Assisted Survey Sampling*
- Wayne A. Fuller, *Sampling Statistics*

Missing data

- Roderick Little, Donald Rubin, *Statistical Analysis with Missing Data*
- Anastasios A. Tsiatis, *Semiparametric Theory and Missing Data*

Causal inference

- Miguel A. Hernán & James M. Robins, *Causal Inference: What If*
- Ding Peng, *A First Course in Causal Inference*
- Judea Pearl, *Causality*

5 Course Grades

- Homework (7 total): **70%**
- Final presentation: **30%**

Homework: Assigned every two weeks; each set contains 3–5 problems.

Final presentation: Each student (or team) will deliver a short presentation on a paper related to survey sampling, missing data, or causal inference. The presentation should clearly state the problem formulation, summarize the paper's main contributions, and highlight the key ideas and insights.

6 Tentative Schedule

Wk	Contents
1	• Course roadmap; *what is statistics* and *what is statistical inference*; data-collection mechanisms as first-class objects (design / missingness / assignment). • Review: LLN/CLT, the delta method, and the intuition behind weighting.
2	• Survey sampling I: finite-population inference; SRSWOR/SRSWR. • Horvitz–Thompson estimator; variance under SRS; confidence intervals.

Wk	Contents
3	• Survey sampling II: unequal-probability designs and PPS sampling. • Stratified and cluster sampling: allocation and efficiency; design effects; two-stage designs.
4	• Survey sampling III: Estimation under complex designs. • Ratio estimators and regression/GREG; design-based vs. model-assisted perspectives. • Variance estimation: Taylor linearization and an influence-function viewpoint.
5	• Survey sampling IV: practical issues and advanced topics; introduction to nonresponse. • Nonresponse adjustment: unit vs. item nonresponse; response mechanisms and bias. • Post-stratification and calibration weighting; response propensity modeling.
6	• Missing data I: MCAR/MAR/MNAR; observed-data likelihood. • Identifiability under MAR; hypothesis testing.
7	• Missing data II: likelihood-based inference for incomplete data. • EM algorithm; selection vs. pattern-mixture models.
8	• Missing data III: multiple imputation (MI) and Rubin's rules. • Practical imputation strategies and post-imputation analysis.
9	• Missing data IV: IPW under MAR; augmented IPW and doubly robust estimation. • Cross-fitting intuition and why overfitting can invalidate inference.
10	• Missing data V: Semiparametric theory.
11	• Causal inference I: potential outcomes; estimands (ATE/ATT/CATE). • Randomized experiments; regression adjustment; precision gains.
12	• Causal inference II: observational studies; unconfoundedness and overlap. • Propensity-score methods: matching, stratification, weighting.

Wk	Contents
13	• Causal inference III: IPW instability; AIPW/DR estimators. • Influence-function intuition; cross-fitting / DML-style inference.
14	• Causal inference IV: outcome modeling vs. weighting; the targeting principle and an introduction to TMLE-style ideas. • Brief overview of modern learners.
15	• Causal inference V: beyond unconfoundedness—instrumental variables (LATE, 2SLS) and difference-in-differences (DiD). • Sensitivity analysis for hidden confounding (overview).
16	• Course wrap-up and final presentations.

Note on the tentative schedule: The schedule is *flexible* and may be adjusted as the course progresses. We will occasionally include *open, high-level discussions* on loosely related statistical ideas—with an open mind and a free exchange of perspectives—to strengthen intuition and cultivate statistical thinking. In addition, we will discuss *selected recent papers* throughout the semester to connect classical theory to modern methodology and current research practice.