BSTA 7820
Statistical Methods for Incomplete Data
Spring, 2026
Anat-Chem 202

Instructor:	Qi Long, 3W 100B, 3600 CCB, (215)573-0659 qlong@upenn.edu
Lectures:	January 14 at 10:15am-1:15pm January 26 – April 27, Monday at 10:15am-1:15pm Canceled lectures: March 9 (Spring Break); March 16 (ENAR 2026 Spring Meeting); April 13
Office Hours:	3W 100B, 3600 Civic Center Boulevard (CCB), Monday 4:30–5pm (cancelled when lecture is cancelled) or by appointment
Textbooks:	Statistical Analysis with Missing Data, 3rd Edition, by Little, R.J.A, and Rubin, D., John Wiley & Sons (2019). Semiparametric Theory and Missing Data, by Tsiatis, A.A., Springer (2007).
Prerequisites:	BSTA 621/622, BSTA 632, BSTA 651, or their equivalents; permission of instructor. Knowledge about Bayesian modeling, though not required, can be helpful.
Course Description:	This course reviews the theory and methodology of incomplete data, covering missing data patterns, missing data mechanisms including MCAR, MAR, and MNAR, potential impacts of missing data on data analysis; imputation methods; likelihood-based methods for handling missing data; computational methods such as the EM algorithm and its extensions; semiparametric methods for missing data such as IPW and AIPW; methods for MNAR and nonignorable missingness including sensitivity analysis. If time permits, it will also cover additional advanced topics on analysis of incomplete data

Outline of Lectures

- Part 1: Introduction (missing data patterns; missing data mechanisms; overview of missing data methods).
- Part 2: Ad hoc methods for handling missing data (complete-case analysis; available-case analysis; LOCF).
- Part 3: Single and multiple imputation methods.
- Part 4: Likelihood-based methods; EM algorithm.
- Part 5: Inverse Probability Weighting (IPW) and Augmented IPW (AIPW) methods.
- Part 6: Methods for handling Missing Not At Random (MNAR) including pattern mixture models, selection models, and sensitivity analysis.
- Part 7 (if time permits): Advanced topics for analysis of incomplete data, e.g., fairness in analysis of incomplete data, and machine learning (ML) and deep learning (DL) imputation models for incomplete high-dimensional data etc.

Grading Policy:

- Attendance and Participation @ 30%
- Homework @ 50%: 4 homework assignments with 12.5% for each
- Final presentation @ 20%

Grades:

- $(85, 100] \approx A$
- $(75, 85] \approx B$
- $(59, 75] \approx C$
- +/- grades will be given accordingly.

Final Presentation: The final presentation will entail a review of 1-2 papers related to analysis of incomplete data and is scheduled for April 27 at 10:15am-1:15pm in TBD. Each student will sign up for a 15-min slot excluding Q&A. Students are expected to provide the instructor by April 6 the list of paper(s) that will be presented.

Misc Notice: All course materials (e.g., outlines, handouts, syllabus exams, PowerPoint presentations, lectures, audio and video recordings, et.c) are proprietary. Students are prohibited from posting, sharing, or selling any such course materials without the express written permission of the professor teaching this course.