University of Maryland School of Public Health

EPIB 664 – Missing Data Analysis

Semester: Fall 2019

Section: 0101

Classroom and Time: Monday 1:00-3:45pm, SPH 0303

Course webpage: *https://umd.instructure.com/courses/1271087*

Instructor: Charles Ma, Ph.D.

Office Hours: Monday 11:00-12:00pm or by appointment

Office: 2234M

Phone: 301-405-6421

Email: tma0929@umd.edu

Course Description: Missing data is a common problem in almost all scientific fields, ranging from biomedical science to social science. Data can be missing unexpectedly during collection process or deliberately by design. Carefully handling the missingness will help us make inference about the population targeted by the complete data. In this course, students will learn the different patterns and mechanisms of missing data, common procedures to handle missingness including imputation-based procedure, likelihood-based procedure and weighting procedure. Useful and popular imputation methods and tools will be introduced. Numerous real data examples will be included to help students understand and solve the real world problem with missing data.

Course Prerequisites:

Required: EPIB 650 *Biostatistics I* and EPIB 651 *Biostatistics II* or permission of instructor.

Recommended: Previous experience with at least one statistical software (e.g. SAS*, R*, STATA)
*: SAS and R are the main softwares used for demonstration in class. Basic introduction and implementation of R will be covered in this course.

Course Learning Objectives:
Upon completing this course, the student will be able to:

1. Understand what is a missing data problem, distinguish the different patterns and mechanisms of missing data.
2. Understand the difference among missing data, complete data and imputed data.
3. Choose the appropriate procedures to handle missingness for different scenarios.
4. Choose the appropriate imputation procedures to impute the missing data when necessary.
5. Use statistical software and tools to conduct an appropriate missing data analysis.
6. Check assumptions of the different missing data models.
7. Interpret the missing data analysis results and know how to make inference.

Program Competencies Addressed in this Course:
The following competencies for the *Master of Public Health with concentration in Biostatistics* are addressed in this course:

1. Distinguish among the different measurement scales or types of variables and select appropriate descriptive statistical methods for summarizing public health data.

2. Select appropriate inferential statistical methods to answer research questions relevant to public health research.
3. Conduct descriptive and inferential statistical analyses that are appropriate to different basic study designs used in public health research.
4. Critically evaluate statistical analyses presented in public health literature.
5. Use statistical analytical software packages (e.g. SAS, R, STATA) to describe, explore, and summarize data as well as perform statistical procedures.
6. Communicate results of statistical analyses to lay and professional audiences.
7. Analyze quantitative data using biostatistics, informatics, computer-based programming and software, as appropriate.
8. Interpret results of data analysis for public health research, policy or practice.

Recommended Texts and Other Readings:

Main reference book:

Enders, C.K. Applied missing data analysis. Guilford press, 2010. **(Enders)**

Other reference book:

Little, R. J. & Rubin, D. B. (2002). *Statistical analysis with missing data (Vol. 333)*. John Wiley & Sons. 2nd edition. **(LR2)**

Websites:

Companion website for the book "Applied missing data analysis": http://www.appliedmissingdata.com/
NC State course "Statistical Methods for Analysis With Missing Data": https://www4.stat.ncsu.edu/~davidian/st790/
SAS PROC MI at UCLA IDRE: https://stats.idre.ucla.edu/sas/seminars/multiple-imputation-in-sas/mi_new_1/
R "mice" package: https://cran.r-project.org/web/packages/mice/index.html

Resources to learn R:

Wickham, H. (2014). *Advanced R*. Chapman and Hall/CRC. (http://adv-r.had.co.nz/)
Crawley, M. J. (2012). *The R book*. John Wiley & Sons.
R manuals: https://cran.r-project.org/manuals.html
Quick-R: http://www.statmethods.net
R-bloggers: https://www.r-bloggers.com/
Swirl (learn R, in R): https://swirlstats.com/
R style guide: https://style.tidyverse.org/

Course Requirements:

Homework:
There will be two homework assignments in this class, each assignment you are given two weeks to finish and will be due at the beginning of the due date class. Electronic copy is allowed. Remember to include your code in the homework you turned in. You are encouraged to discuss with your classmates or work in teams for the homework, but each student has to submit their own homework. **Late homework will be penalized (see Late assignment policy below).**

Exam:
There will one take-home midterm exam. You are given one week to finish the exam. The format will be similar to the homework, however, **students CANNOT discuss midterm with their classmates**. **Late submission will be penalized (see Late assignment policy below).**

Project:
The project in this course accounts for a significant portion (50%) of the grade and should represent student's understanding of missing data analysis. Each student will be responsible for picking one of the missing data projects provided by the instructor or finding a study with missing data to work on, identifying the missing pattern and mechanism, clearly stating the purpose of the analysis, carrying out the appropriate procedure to handle missingness, **writing a report**, and **giving an in-class presentation (~10 mins)**. Grading will be based on both the written report and the oral presentation. More details of the project will be posted later in class.

There is no specific format requirement or page limit for the report. Generally, a formal report should at least include a title, the background introduction, the method applied and the major analysis results. The following outline can be used as a guideline for your report:

- Background: introduce the research area, objective of the study, the hypothesis question, etc.
- Data: describe the data, summary table and some descriptive statistics, the missing pattern/structure and possible missing mechanism (and think about why)
- Statistical methods: what methods (can have more than one) used to address the question of interest and handle missingness (justify your choice)
- Results: report the results of your analysis and include some discussion, possibly compare your results to a complete-case analysis.
- Conclusion: what are the main findings? What are some limitations, etc.

Course Website:
Course announcements, lecture notes, data sets, programs, and homework assignments will be distributed on the course webpage (*https://umd.instructure.com/courses/1271087*). Please check it on a regular basis. Lecture notes will be posted before class. You may wish to print these notes prior to each lecture and use them as an outline for taking notes during the class.

University Course Related Policies:

All University of Maryland-approved graduate course policies are provided here:
https://gradschool.umd.edu/course-related-policies

Critical University Policies:

Inclement Weather / University Closings / Emergency Procedures:
In the event that the University has a delayed opening or is closed for an emergency or extended period of time, the instructor will communicate to students regarding schedule adjustments, including rescheduling of examinations and assignments due to inclement weather and campus emergencies.

Religious Observances:
The University System of Maryland policy provides that students should not be penalized because of observances of their religious beliefs; students shall be given an opportunity, whenever feasible, to make up within a reasonable time any academic assignment that is missed due to individual participation in religious observances.

Special Accommodations / Disability Support Services:
If you have a documented disability and wish to discuss academic accommodations for test taking or other needs, you will need documentation from Disability Support Service (301-314-7682). If you are ill or encountering personal difficulties, please let the instructor know as soon as possible. You can also contact Learning Assistance Services (301-314-7693) and/or the Counseling Center (301-314-7651) for assistance.

Academic Integrity:
The University's code of academic integrity is designed to ensure that the principle of academic honesty is upheld. Any of the following acts, when committed by a student, constitutes academic dishonesty:

- CHEATING: intentionally using or attempting to use unauthorized materials, information, or study aids in an academic exercise.
- FABRICATION: intentional and unauthorized falsification or invention of any information or citation in an academic exercise.
- FACILITATING ACADEMIC DISHONESTY: intentionally or knowingly helping or attempting to help another to violate any provision of this code.
- PLAGIARISM: intentionally or knowingly representing the words or ideas of another as one's own in any academic exercise.
-

For more information see: *http://www.shc.umd.edu/code.html*.

The Honor Pledge is a statement undergraduate and graduate students should be asked to write by hand and sign on examinations, papers, or other academic assignments. The Pledge reads:
I pledge on my honor that I have not given or received any unauthorized assistance on this assignment/examination.
The University of Maryland, College Park has a nationally recognized Code of Academic Integrity, administered by the Student Honor Council. This Code sets standards for academic integrity at Maryland for all undergraduate and graduate students. As a student you are responsible for upholding these standards for this course. It is very important for you to be aware of the consequences of cheating, fabrication, facilitation, and plagiarism. For more information on the Code of Academic Integrity or the Student Honor Council, please visit *http://www.shc.umd.edu/* .

Course Policies:

Late Assignment Policy:
Full credit will be given for assignments turned in on the due date. The assignment should be turned in before class. 80% credit for one day late, 50% credit for two days late. NO credit given after two days late. If sickness or emergency, no deduction will be taken if the lecturer is informed before the homework/midterm is due.

Grading Procedures:

Grade of this course will be determined as follows:

Homework	30%
Midterm Exam	20%
Project	50%

Course Outline / Course Calendar:

Tentative Course Schedule*			
Session	**Date**	**Topic**	**Assignments****
1	8/26/2019	Introduction to missing data: basic concepts, missing data pattern and missing mechanism	
-	9/2/2019	Labor day, No Class	
2	9/9/2019	Introduction to R I	
3	9/16/2019	Introduction to R II	HW1 assigned
4	9/23/2019	Traditional methods to handle missingness: deletion methods and single imputation	
5	9/30/2019	Likelihood-based methods I: maximum likelihood inference for full data and missing data	HW1 due
6	10/7/2019	Likelihood-based methods II: Expectation-Maximization (EM) algorithm, improving the accuracy of maximum likelihood analyses	HW2 assigned
7	10/14/2019	Multiple Imputation I: introduction to Bayesian estimation and imputation phase	
8	10/21/2019	Multiple Imputation II: analysis and pooling phase, practical issue	HW2 due
9	10/28/2019	Multiple Imputation III: software	
-	11/4/2019	APHA meeting, No Class	Midterm assigned
10	11/11/2019	Weighting procedure to handle missingness	Midterm due
11	11/18/2019	Nonignorable missing data models	
12	11/25/2019	Application of missing data analysis methods in real world studies***	
13	12/2/2019	Student in-class presentation	Project due

* This is a tentative schedule, and the actual materials covered in each lecture might not be exactly the same.

** Homework assigned and due dates might be subject to change.

*** Depending on the class size, part of this class may be used for student presentation.

Note: Numbers in brackets after learning objectives show linkage between material covered in each session and the numbered program competencies shown on page 1 of this syllabus.

Required Session Outline

Session 1 **8/26/2019**

Topic: Introduction to missing data: basic concepts, missing data pattern and missing mechanism

Learning Objectives for Session 1 [Relevant Program Competencies: #1, #2, #3, #4]

- Distinguish among the different measurement scales or types of variables and select appropriate descriptive statistical methods for summarizing public health data.
- Select appropriate inferential statistical methods to answer research questions relevant to public health research.
- Critically evaluate statistical analyses presented in public health literature.

Reading: Enders Chapter 1, LR2 Chapter 1, 2

Session 2 **9/9/2019**

Topic: Introduction to R I

Learning Objectives for Session 2 [Relevant Program Competencies: #5, #7]

- Use statistical analytical software packages (e.g. SAS, R, STATA) to describe, explore, and summarize data as well as perform statistical procedures.
- Analyze quantitative data using biostatistics, informatics, computer-based programming and software, as appropriate.

Session 3 **9/16/2019**

Topic: Introduction to R II

Learning Objectives for Session 2 [Relevant Program Competencies: #5, #7]

- Use statistical analytical software packages (e.g. SAS, R, STATA) to describe, explore, and summarize data as well as perform statistical procedures.

Analyze quantitative data using biostatistics, informatics, computer-based programming and software, as appropriate.

Session 4 **9/23/2019**

Topic: Traditional methods to handle missingness: deletion methods and single imputation

Learning Objectives for Session 4 [Relevant Program Competencies: #2, #3, #5, #6, #7, #8]

- Select appropriate inferential statistical methods to answer research questions relevant to public health research.
- Conduct descriptive and inferential statistical analyses that are appropriate to different basic study designs used in public health research.
- Use statistical analytical software packages (e.g. SAS, R, STATA) to describe, explore, and summarize data as well as perform statistical procedures.
- Communicate results of statistical analyses to lay and professional audiences.
- Analyze quantitative data using biostatistics, informatics, computer-based programming and software, as appropriate.
- Interpret results of data analysis for public health research, policy or practice.

Reading: Enders Chapter 2, LR2 Chapter 3, 4

Session 5 **9/30/2019**

Topic: Likelihood-based methods I: maximum likelihood inference for full data and missing data

Learning Objectives for Session 5 [Relevant Program Competencies: #2, #3, #5, #6, #7, #8]

- Select appropriate inferential statistical methods to answer research questions relevant to public health research.
- Conduct descriptive and inferential statistical analyses that are appropriate to different basic study designs used in public health research.
- Use statistical analytical software packages (e.g. SAS, R, STATA) to describe, explore, and summarize data as well as perform statistical procedures.
- Communicate results of statistical analyses to lay and professional audiences.
- Analyze quantitative data using biostatistics, informatics, computer-based programming and software, as appropriate.
- Interpret results of data analysis for public health research, policy or practice.

Reading: Enders Chapter 3, 4

Session 6 **10/7/2019**

Topic: Likelihood-based methods II: Expectation-Maximization (EM) algorithm, improving the accuracy of maximum likelihood analyses

Learning Objectives for Session 6 [Relevant Program Competencies: #2, #3, #5, #6, #7, #8]

- Select appropriate inferential statistical methods to answer research questions relevant to public health research.
- Conduct descriptive and inferential statistical analyses that are appropriate to different basic study designs used in public health research.
- Use statistical analytical software packages (e.g. SAS, R, STATA) to describe, explore, and summarize data as well as perform statistical procedures.
- Communicate results of statistical analyses to lay and professional audiences.
- Analyze quantitative data using biostatistics, informatics, computer-based programming and software, as appropriate.
- Interpret results of data analysis for public health research, policy or practice.

Reading: Enders Chapter 4, 5

Session 7 **10/14/2019**

Topic: Multiple Imputation I: introduction to Bayesian estimation and imputation phase

Learning Objectives for Session 7 [Relevant Program Competencies: #2, #3, #5, #6, #7, #8]

- Select appropriate inferential statistical methods to answer research questions relevant to public health research.
- Conduct descriptive and inferential statistical analyses that are appropriate to different basic study designs used in public health research.
- Use statistical analytical software packages (e.g. SAS, R, STATA) to describe, explore, and summarize data as well as perform statistical procedures.
- Communicate results of statistical analyses to lay and professional audiences.
- Analyze quantitative data using biostatistics, informatics, computer-based programming and software, as appropriate.
- Interpret results of data analysis for public health research, policy or practice.

Reading: Enders Chapter 6, 7

Session 8 **10/21/2019**

Topic: Multiple Imputation II: analysis and pooling phase, practical issue

Learning Objectives for Session 8 [Relevant Program Competencies: #2, #3, #5, #6, #7, #8]

- Select appropriate inferential statistical methods to answer research questions relevant to public health research.
- Conduct descriptive and inferential statistical analyses that are appropriate to different basic study designs used in public health research.
- Use statistical analytical software packages (e.g. SAS, R, STATA) to describe, explore, and summarize data as well as perform statistical procedures.
- Communicate results of statistical analyses to lay and professional audiences.
- Analyze quantitative data using biostatistics, informatics, computer-based programming and software, as appropriate.
- Interpret results of data analysis for public health research, policy or practice.

Reading: Enders Chapter 8, 9

Session 9 **10/28/2019**

Topic: Multiple Imputation III: software

Learning Objectives for Session 9 [Relevant Program Competencies: #5, #6, #7, #8]

- Use statistical analytical software packages (e.g. SAS, R, STATA) to describe, explore, and summarize data as well as perform statistical procedures.
- Communicate results of statistical analyses to lay and professional audiences.
- Analyze quantitative data using biostatistics, informatics, computer-based programming and software, as appropriate.
- Interpret results of data analysis for public health research, policy or practice.

Reading: Enders Chapter 11

Session 10 **11/11/2019**

Topic: Weighting procedure to handle missingness

Learning Objectives for Session 10 [Relevant Program Competencies: #2, #3, #5, #6, #7, #8]

- Select appropriate inferential statistical methods to answer research questions relevant to public health research.
- Conduct descriptive and inferential statistical analyses that are appropriate to different basic study designs used in public health research.
- Use statistical analytical software packages (e.g. SAS, R, STATA) to describe, explore, and summarize data as well as perform statistical procedures.
- Communicate results of statistical analyses to lay and professional audiences.
- Analyze quantitative and qualitative data using biostatistics, informatics, computer-based programming and software, as appropriate.
- Interpret results of data analysis for public health research, policy or practice.

Session 11 **11/18/2019**

Topic: Nonignorable missing data models

Learning Objectives for Session 11 [Relevant Program Competencies: #2, #3, #5, #6, #7, #8]

- Select appropriate inferential statistical methods to answer research questions relevant to public health research.
- Conduct descriptive and inferential statistical analyses that are appropriate to different basic study designs used in public health research.
- Use statistical analytical software packages (e.g. SAS, R, STATA) to describe, explore, and summarize data as well as perform statistical procedures.
- Communicate results of statistical analyses to lay and professional audiences.
- Analyze quantitative data using biostatistics, informatics, computer-based programming and software, as appropriate.
- Interpret results of data analysis for public health research, policy or practice.

Reading: Enders Chapter 10

Session 12 **11/25/2019**

Topic: Application of missing data analysis methods in real world studies

Learning Objectives for Session 12 [Relevant Program Competencies: #3, #5, #6, #7, #8]

- Conduct descriptive and inferential statistical analyses that are appropriate to different basic study designs used in public health research.
- Use statistical analytical software packages (e.g. SAS, R, STATA) to describe, explore, and summarize data as well as perform statistical procedures.
- Communicate results of statistical analyses to lay and professional audiences.
- Analyze quantitative data using biostatistics, informatics, computer-based programming and software, as appropriate.
- Interpret results of data analysis for public health research, policy or practice.

Session 13 **12/2/2019**

Topic: Student in-class presentation

Learning Objectives for Session [Relevant Program Competencies: #1, #2, #3, #4, #5, #6, #7, #8]

- Distinguish among the different measurement scales or types of variables and select appropriate descriptive statistical methods for summarizing public health data.
- Select appropriate inferential statistical methods to answer research questions relevant to public health research.
- Conduct descriptive and inferential statistical analyses that are appropriate to different basic study designs used in public health research.
- Critically evaluate statistical analyses presented in public health literature.
- Use statistical analytical software packages (e.g. SAS, R, STATA) to describe, explore, and summarize data as well as perform statistical procedures.
- Communicate results of statistical analyses to lay and professional audiences.
- Analyze quantitative data using biostatistics, informatics, computer-based programming and software, as appropriate.
- Interpret results of data analysis for public health research, policy or practice.