Lecture 7
Missing Data

Instructor: Michela Taufer

THE UNIVERSITY OF
TENNESSEE
KNOXVILLE
BIG ORANGE. BIG IDEAS.®

Today we will explore …

- Reasons for missing data
- Strategies for dealing with datasets with missing data
- Apply the strategies to real dietary data

Data Collection

Missing data = incomplete observations

- Critical data issues:
  - Reasons for missing data
  - Scale and distribution of the values in the data

Case Study

- Study of an asthma education intervention in eight schools
- Randomly chosen set of students aged 8 to 14 with asthma
- Observations over two weeks period post-treatment
- Students complete:
  - Scale to measure self-efficacy beliefs with regard to their asthma
  - Questionnaire rating severity of their symptoms

(Velsor-Friedrich, see attached paper in GitHub)

Case Study

- Students simply forgot to visit the school clinic to fill out the form → Missing completely at random (MCAR)
  - Complete cases are representative of the originally sample
- Students missed school because of severity of their asthma symptoms and failed to complete the symptom severity rating
  - Missing variable is directly related to study
  - Example of non-ignorable missing data!!!
- Younger children missed ratings of symptom severity because they had a harder time interpreting the rating form → missing at random (MAR)
  - Values are missing for reasons related to another observable variable

Mechanisms to Deal with Missing Data

- Data are MCAR or MAR
  - Ignore the reasons for missing data in the analysis of the data
  - Simplify the model-based methods used for missing data analysis
- Use more than one method for collecting important information
  - E.g., income + years of education or type of employment

Table 1. Variable Descriptions.

Variable	Definition	Possible values	*M*	(*SD*)	*N*
Asthma belief Survey	Level of confidence in controlling asthma	Range from 1, little confidence to 5, lots of confidence	4.057	(0.713)	154
Group	Treatment or control group	0 = Treatment 1 = Control	0.558	(0.498)	154
Symsev	Severity of asthma symptoms in 2 week period post-treatment	0 = no symptoms 1 = mild symptoms 2 = moderate symptoms 3 = severe symptoms	0.235	(0.370)	141
Reading	Standardized state reading test score	Grade equivalent scores, ranging from 1.10 to 8.10	3.443	(1.636)	79
Age	Age of child in years	Range from 8 to 14	10.586	(1.605)	152
Gender	Gender of child	0 = Male 1 = Female	0.442	(0.498)	154
Allergy	Number of allergies reported	Range from 0 to 7	2.783	(1.919)	83

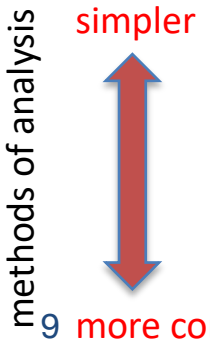
Table 1. Variable Descriptions.

POPULATION: 154

Variable	Definition	Possible values	*M*	(*SD*)	*N*
Asthma belief Survey	Level of confidence in controlling asthma	Range from 1, little confidence to 5, lots of confidence	4.057	(0.713)	154
Group	Treatment or control group	0 = Treatment 1 = Control	0.558	(0.498)	154
Symsev	Severity of asthma symptoms in 2 week period post-treatment	0 = no symptoms 1 = mild symptoms 2 = moderate symptoms 3 = severe symptoms	0.235	(0.370)	141
Reading	Standardized state reading test score	Grade equivalent scores, ranging from 1.10 to 8.10	3.443	(1.636)	79
Age	Age of child in years	Range from 8 to 14	10.586	(1.605)	152
Gender	Gender of child	0 = Male 1 = Female	0.442	(0.498)	154
Allergy	Number of allergies reported	Range from 0 to 7	2.783	(1.919)	83

Table 2. Missing Data Patterns.

Symsev	Reading	Age	Allergy	# of cases	% of cases
O	O	O	O	19	12.3
M	O	O	O	1	0.6
O	M	O	O	54	35.1
O	O	O	M	56	36.4
M	M	O	O	9	5.8
M	O	O	M	1	0.6
O	M	O	M	10	6.5
O	O	M	M	2	1.3
M	M	O	M	2	1.3
# missing	# missing	# missing	# missing	154	
13 (8.4%)	75 (48.7%)	2 (1.3%)	71 (46.1)		



What is the reasons for the missing data?

Can we accept the MCAR assumption?

Can we accept the MAR assumption?

Does missing data result from a non-ignorable response mechanism?

Commonly-Used Missing Data Methods

- Complete-Case Analysis: Cases that are missing variables in the proposed model are dropped from the analysis, leaving only complete cases
  - Assume that missing data are MCAR
  - Adequate amount of data remains for the analysis?

Commonly-Used Missing Data Methods

- Available Case Analysis: with X1 complete and X2 partially complete, all cases are used to estimate the mean of X1, but only the complete cases contribute to an estimate of X2, and the correlation between X1 and X2.
  - Different sets of cases are used to estimate parameters of interest in the data

Strategy 1

X1 = x11 x12 x13 x14 x15 → work on a population of 5 individuals

X2 = x22 x34 x25 → work on a population of 3 individuals

Strategy 2

X1 = x12 x14 x15 → work on a population of 3 individuals

X2 = x22 x34 x25 → work on a population of 3 individuals

Commonly-Used Missing Data Methods

- Single-Value Imputation: Fill in the missing value with a plausible one, e.g., mean for cases that observe the variable
  - Analyst continues with the statistical method as if the data are completely observed
  - Single value changes the distribution of that variable by decreasing the variance that is likely present
  - Bias in the estimation of variances and standard errors are compounded

Strategy 3

X1 = x11 x12 x13 x14 x15 → work on a population of 5 individuals

X2 = x_{avg} x22 x_{avg} x34 x25 → work on a population of 5 individuals

Model-Based Methods

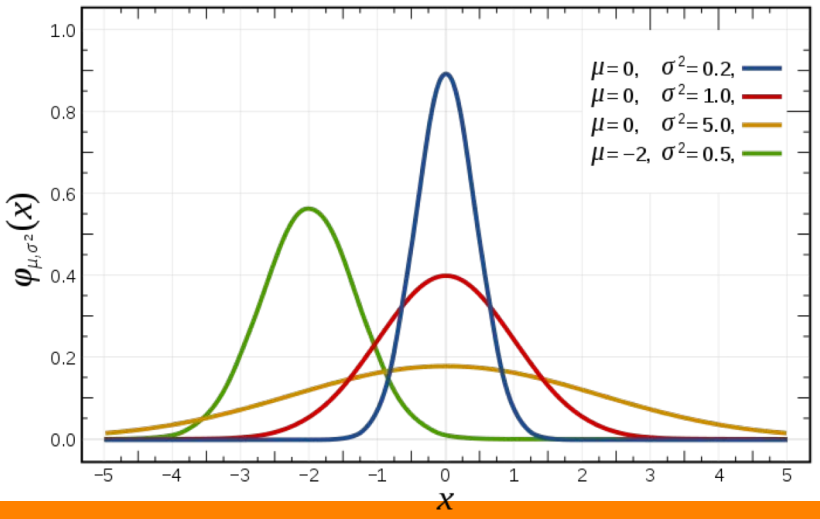
- Add assumptions about the distribution of the data and the nature of the missing data mechanism
  - Multiple imputation

Strategy 4

X1 = x11 x12 x13 x14 x15 → work on a population of 5 individuals

X2 = x_{21} x22 x_{23} x34 x25 → work on a population of 5 individuals

Assumption:



Relevant Open-source Dataset

- Use a dataset with well-known and broadly used data format

NHANES: National Health and Nutrition Examination Survey

  - Medical, demographic, and dietary records
  - Available to the public for free
  - *Contains subjective food groups provided by USDA*

Data available at: http://www.cdc.gov/nchs/nhanes/nhanes_questionnaires.htm

NHANES Dietary Data

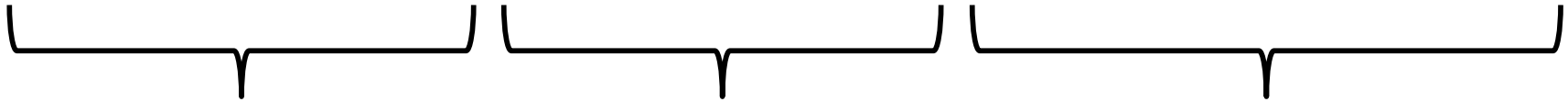
- Dietary intake of 64,653 Americans
- 7,494 unique food items
- 1,587,750 food entries
- 46 nutrient features for each food item
  - Macronutrients (e.g., fats, carbohydrates)
  - Micronutrients (e.g., vitamins, minerals)

NHANES Dietary Data

- Dietary intake of 64,653 Americans
- **7,462 unique food items**
- 1,587,750 food entries
- 46 nutrient features for each food item
  - **Macronutrients (e.g., fats, carbohydrates, proteins)**
  - Micronutrients (e.g., vitamins, minerals)

Structure of Dietary Data Item

<143672, 92510000, 3, 0, 8:15am, 7, 3, 10.1, 4, 3.45, 10, 178, …>



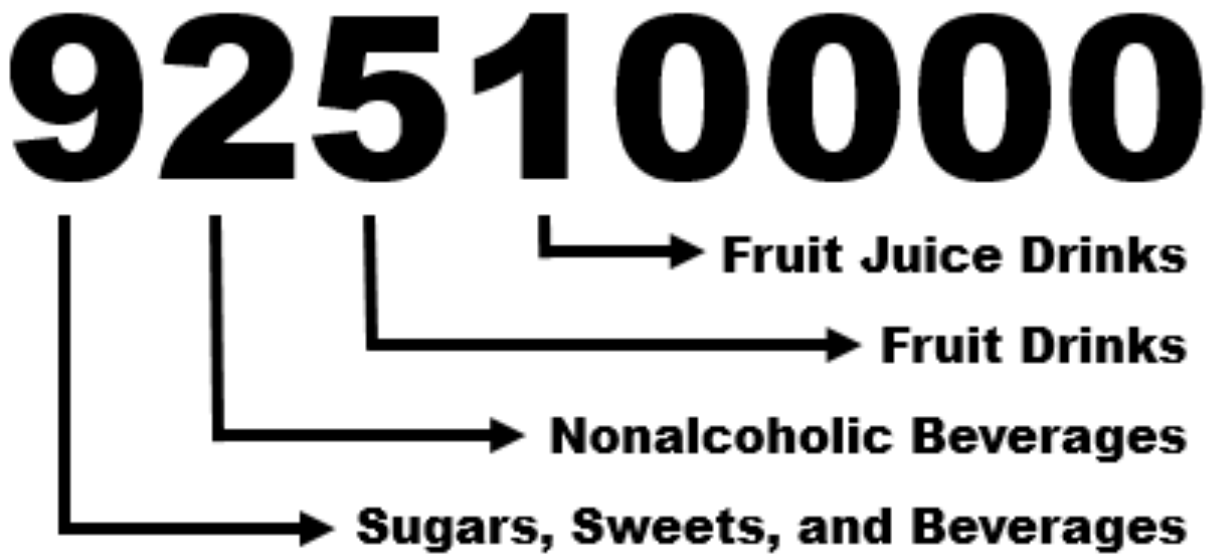
- Participant ID
- **USDA Food Code**

- Meta Data

- **Macronutrients**
- Micronutrients

USDA Food Classification

- Subjective and general
- Categorical, not nutrient-driven



Assignment 7 – Problem 1

- Data file: ./data/data-1.csv
  - Contain no missing values
- Problem: Cluster food items in the file based on carbohydrate and fat:
  - *Use the k-Means from the Spark MLlib to cluster data points*
  - *Determine the optimal value for k using the* **elbow method**
- Metric of quality: Use Within Set Sum of Squared Errors
  - This method is built into the Spark kMeans model and can be accessed with model.computeCost()
- Note: the clusters provide a ground truth for comparison to clusters we find later using data with missing values.

Assignment 7 – Problem 1

Steps:

- Define the optimal value for K
- Cluster food items (using K)
- Plot clusters by *fat* and *carbohydrate* content

Assignment 7 – Problem 2

- Data file: ./data/data-2.csv
  - Missing values for the carbohydrate content of some food items
  - The data were removed from a specific set of food items (i.e., food items with carbohydrate value near 0.5)
- Define a method to remove food items with missing any macronutrient values and apply this method to the data.
- Cluster the modified data and plot the results
  - Use the same K value as in Problem 1
- Note: we provide you with the code that loads the data and reports the percentage of values missing for each macronutrient (i.e., carbohydrates and fat)

Assignment 7 – Problem 3

- Data file: ./data/data-3.csv
  - Missing values for the fat or carbohydrate content of some food items (but not both for a single food item).
  - The data were removed from a food items randomly

PART 1:

- Define and apply a method to fill missing values with the mean of other values
  - *E.g., for missing values in fat, fill with the mean of fat values that are present*
- *Cluster the modified data and plot the results*
  - Use the same K value as in Problem 1

Assignment 6 – Problem 3

PART 2:

- *Use the code for Problem 2 to remove data with missing values rather than filling the gaps (as you did in PART 1)*
- *Cluster the modified data and plot the results*
  - Use the same K value as in Problem 1

Assignment 6 – Problem 4

- **Observe and describe:** Can you summarize your findings in each problem? Can you compare and contrast the findings across problems? How did each method for dealing with missing data (i.e., remove of filling) change the clustering outcome?
- **Impact of K:** What value did you choose for K in Problems 1-3? You based the selection of your K on the first dataset (i.e., no missing data). Do you expect a different value of K if you had used the elbow method with the second or third dataset? If yes, propose changes to your current solutions.

Assignment 6 – Problem 4

- **Building assumptions on data distributions:** Now look at the plot of clusters in Problem 1. Logically, there cannot be more than 1 gram of (carbohydrate + fat) in 1 gram of food. In your plot this can be seen in the form of a diagonal line from the top-left to bottom-right (where the sum of fat and carbohydrate content is equal to 1). How can you use this information to improve the way you fill missing values? Can you think of other methods to fill missing values? (HINT: logistic regression)

Project (I)

- What are the scientific questions you are answering with your tool/framework? Be as specific as possible.
- What is the tentative title of your project?
- What are the milestones you want to meet from now until Dec 3 when you will present your poster at the poster showcase? Be as specific as possible. Write the date of the deadline and the task(s) you want to achieve for that deadline.

THE UNIVERSITY OF
TENNESSEE
KNOXVILLE

BIG ORANGE. BIG IDEAS.®

BIG ORANGE BIG IDEAS