Data Management and Visualization

Dr. Ashish Kumar Jha

Session 4

Data Management

Agenda

Data Wrangling

Missing Data Management

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Trinity Business School

Data Handling

Cleaning and Managing Data

Infinite universe

Most important task

Understand an play with your data before using it

Could require upto 80% time of project

Cleaning and managing data in R

Major packages

- Dplyr
- Tidyr
- Get acquainted with tidyverse
  - R for datascience

Major activities to undertake

Load data in R

- Read CSV
- Readxl

View data

- Tidyr::glimpse()
- utils::View(iris)

Identify datatype

- Typeof()

Preliminary Analysis

Cleaning and arranging

- Gsub()
- Gather()
- Distinct()

The pipe operation

- %>% glimpse()

Combine files

- List.files()
- Read csv()
- Rbind()

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Trinity Business School

Missing Data

Why Care?

- Missing values are very common
- Difficult to deal with missing data
- Difficult to interpret results with missing data
- The simplest method to deal with missing data is data reduction which deletes the instances with missing values. However it will lead to great information loss

Causes

Random Error

- Data missing due to random error
- Someone forgot some item in a survey
- Some data was missing because of a clerical error

Bias

- Systemic reason for missing data
- Some questions can’t be answered by some people
- Questions like do you some would be missing for smokers

Types of Missing Data

Missing Completely at Random(MCAR)

– certain value is missing has nothing to do with its hypothetical value and with the values of other variables.

Missing at Random (MAR)

– propensity for a data point to be missing is not related to the missing data

Missing not at Random (MNAR)

– missing value depends on the hypothetical value

– missing value is dependent on some other variable’s value

Types of Missing Data

Mathematically Speaking

MCAR

– Probability of missing data is unrelated to missing and observed values

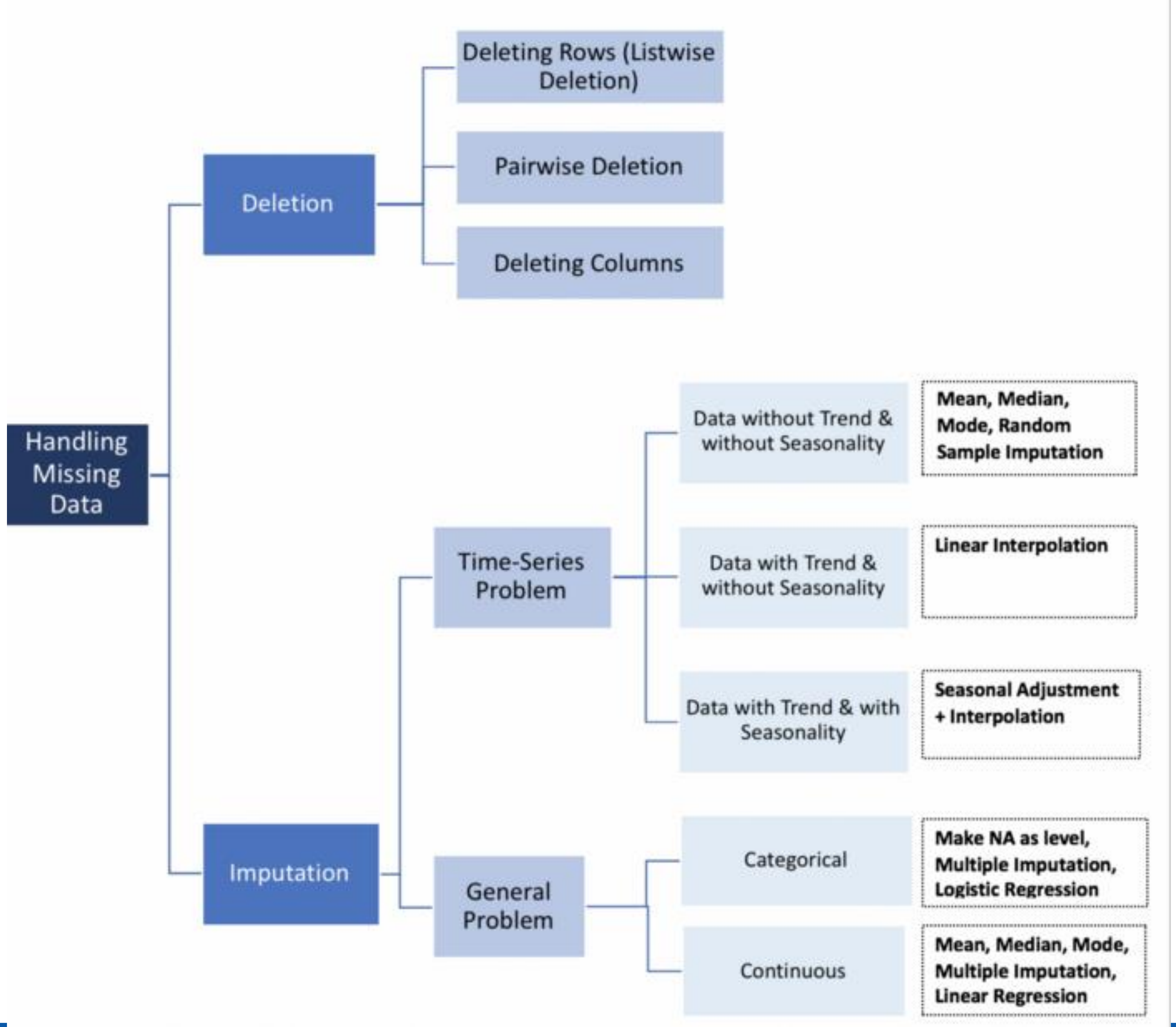
MAR

– Probability of missing data is related to observed values

MNAR

– Probability of missing data is related to unobserved values

Handling Missing Data



Source: https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4

Handling Missing values- Discard/Delete

Discard cases

- Discard rows/columns
- Works on MCAR only
  - High bias if lots of missing data points
  - Could lead to few complete rows of data

Handling Missing Values- Preserve

Mean Substitution

- Popular approach
- replacing the missing values by the mean of all observed values
- Does not work in systemic biases
- Variance decreases

Handling Missing Values- Preserve

Hot Deck Imputation

- replace missing values with a related row.
- Popular in surveys
- Involves replacing missing values of one or more variables for a non-respondent with observed values from a respondent that is similar to the non-respondent with respect to characteristics observed by both cases.

Handling Missing Values- Preserve

Regression- Model based substitution

- Use Linear or logistic regression as appropriate
- In R you can use VIM package
- Popular in large secondary databases
- Could account for some level of biasness- MAR
- Can use time series models as well where appropriate

Handling Missing Values- Preserve

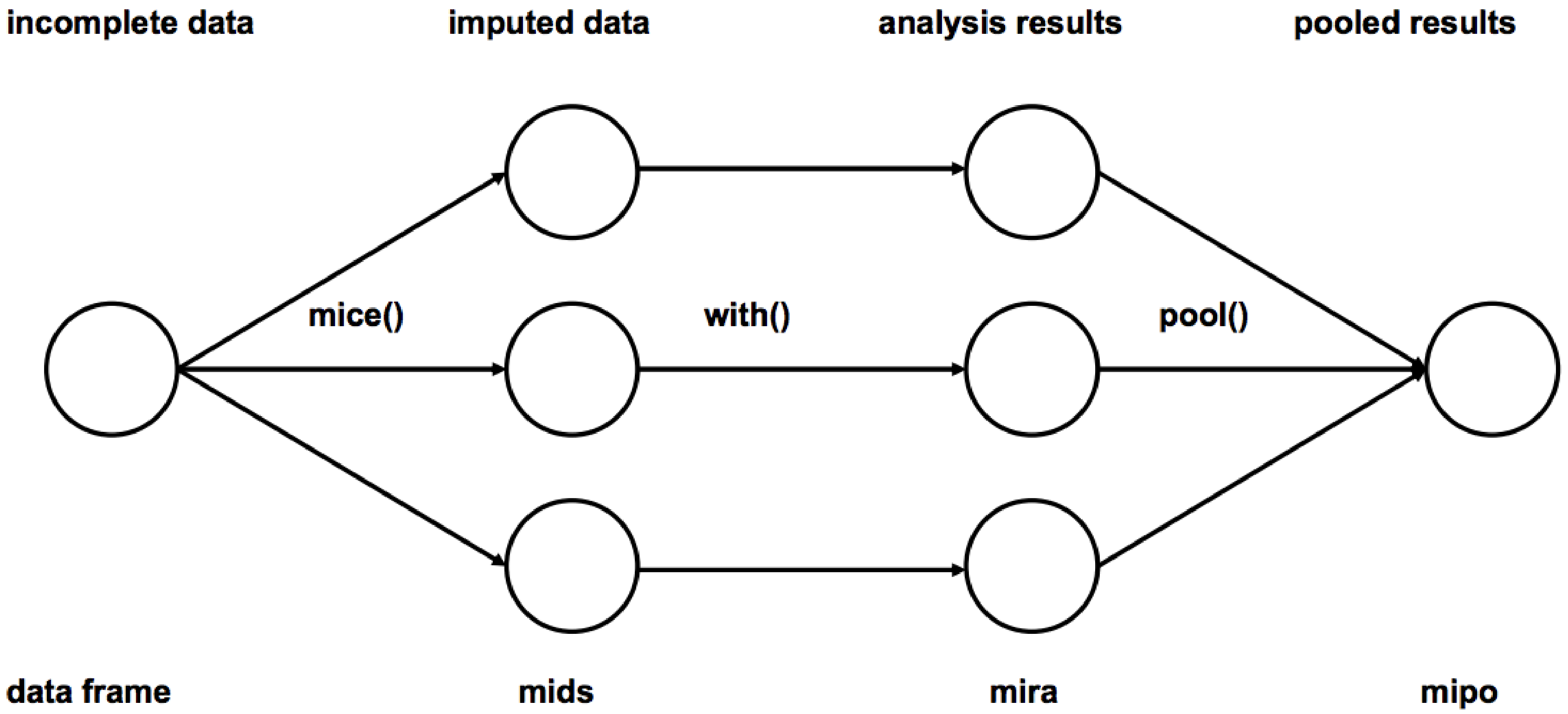
EM Imputation

- EM stands for expectation maximization
- It is an iterative process to calculate the sufficient statistics to impute multiple values
- Popular package Amelia in R
- Expectation-Maximization Bootstrap-based algorithm (EMB) It assumes that the complete data are multivariate normal

Handling Missing Values- Preserve

Multiple Imputation

- One of the most powerful techniques
- Package mi in R



Handling Missing Values- Preserve

Multiple Imputation

- Make a model that predict every missing data item (linear or logistic regression, non-linear models, etc.)
- Use the above models to create a “complete” dataset.
- Each time a “complete” dataset is created, do an analysis of it, keeping the mean and SE of each parameter of interest.
- Repeat this between 2 and tens of thousands of time
- To form final inferences, for each repetition, average across means, and sum the within and between variances for each parameter.