Table des matières

Statistics

Missing Data

Vidéos YouTube

- ritvikmath: Missing Data Mechanisms
- ritvikmath: Dealing With Missing Data Part I
- ritvikmath: Dealing With Missing Data - Multiple Imputation
-

Notes sur les vidéos YouTube

Video 1: ritvikmath: Missing Data Mechanisms

- MCAR
  - Librarians forget to enter the data completely randomly ;
- MAR
  - Women are 90% likely to respond to a survey on the number of overdue books while men are 70% likely to respond ;
- MNAR : **The missingness of a certain value depends on the true value itself**
  - For example :
    - If I have 0 books overdue, I'm 90% likely to respond to a question asking how many overdue books I have ;
    - If I have 1 books overdue, I'm 80% likely to respond to a question asking how many overdue books I have ;
    - If I have 2 books overdue, I'm 70% likely to respond to a question asking how many overdue books I have ;
    - ...
  - So, the more books I have that are overdue, the less likely I am to respond to a question asking how many books I have that are overdue because I may be embarassed or feel shame ;
  - Therefore MNAR is kind of a **chicken and egg** scenario ;
    - If try to figure if a column is MNAR, need to figure out if those missing values are based on the actual values ;

    * But, I *don't know* the **actual** values of that column because they're missing in the first place!
    * So it's really hard to figure out if something is MNAR;
  - In comparision, MCAR and MAR are easier to figure out if something is one or the other;
  - Can slice a dataset by values of a column
    * For example, by sex or by age group, . . . ;
    * If missing value rate is about the same for all different slices then likely to be Missing Completely At Random;
    * If however it's different for each slice, then it's likely MAR;

Video 2: ritvikmath: Dealing With Missing Data Part I

- Row deletion;
  - Most common and easiest;
  - Omit any row in dataset with a missing value—pretend it does not exist;
  - Seems too good to be true because it usually is;
  - Can only do this if the data is Missing Completely At Random—biased otherwise;

    Makes sense if you reason that any other way you'd obviously be creating a bias in your data;

    Each column would have missing values completely at random and without respect to, for example, the gender and the estimations would be *unbiased*;
  - Thus, have to be very careful it's really what we want to do because likely cause bias if there's any sort of relationship between the missing variables and other columns;

Pros	Cons
simple	**biased**

- Mean/Median imputation;

- A little more « clever » ;
- **Seems** intuitive and is pretty simple ;
- Mean
  * Fill in, for example, 1.8 as the average of a few discrete values ;
  * BUT, will **artifically** *reduce variability* of the data ;
  * *Seem* like several values have the exact same value ;
- Median is the same idea but will overepresent one fixed value ;

Pros	Cons
simple	**lower variability**

› Hot Deck methods ;

- Most clever so far ;
- Any family of methods where :
  * Compute a missing value based on the value of examples that are *similar* to it ;
- For example
  * Fill in the missing value of a female by the average of only other females ;
- Better because imputing more information (whether someone is female or male) ;
  * Imagine if there's a bunch of columns (income, family members, where they live, etc.) ;
  * Then, we can input missing values based on a few people that are really similar to the missing person ;
  * Logical as we would expect that person's missing value to be similar to other similar people's ;
- **May not be true**, but it's a very ***educated*** *guess* ;

Pros	Cons
more educated	more (computationnaly) expensive

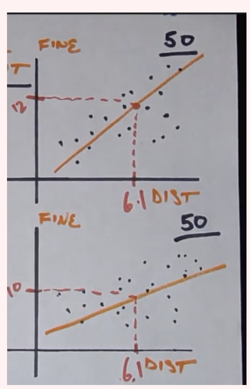
Video 3: ritvikmath: Dealing With Missing Data - Multiple Imputation

- *single* imputation implique qu'on se ramasse avec une seule valeur (peu importe ce que c'est) ;
  - Régression, moyenne, médiane, etc. sont tous une seule valeur.
- *multiple* imputation even more clever than hot-deck methods ;
- For example, regression of library fees in function of kilometer distance from the library
  - Sample 50 data points from thousands, and estimate fees for a given distance ;
  - Repeat with different samples ;

    Generally the more repetitions, the less biased the estimations but 5 is a good rule of thumb ;
  - Treat each predicted value as a complete observation ;
  - Then with this "complete" data set, we do what we want ;
  - Take some kind of aggregate of all the values we wanted from each of 5-ish set ;
  - Analyze how far from each other the aggregated values are ;

    If a lot of variability, bad ;

    If few variability, good-ish because means aggregated values are closer.



- Cons : complicated,