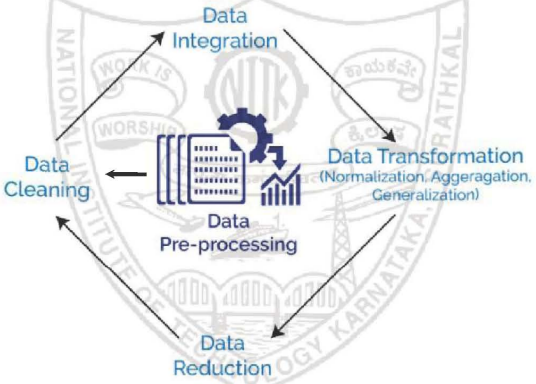
Feb – Jun 2023
IT258

Data Preprocessing

Dr. Sowmya Kamath S, Dept of IT, NITK Surathkal

Knowledge Discovery Process

Major Tasks

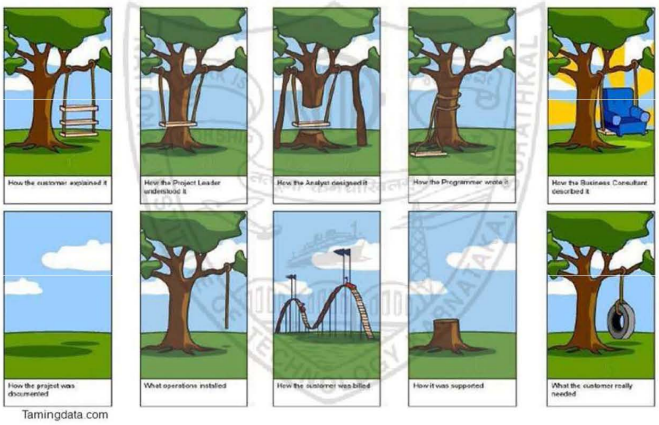


Need for Data Cleaning

No quality data, no quality results!

- Quality decisions can only be based on quality data

Garbage In Garbage Out



Tamingdata.com

Common data issues to be dealt with..

- **Missing values:** how do we estimate/infer, fill in …?
- **Wrong values:** how can we detect and correct?
- **Messy formats:** how to deal with them?
- **Unusable:** if available data cannot answer the question posed, how to manage?

Data Quality

- Accuracy
- Completeness
- Uniqueness
- Timeliness
- Consistency

Data Quality

- Accuracy
  - The data was recorded correctly.
- Completeness
  - All relevant data was recorded.
- Uniqueness
  - Entities are recorded once.
- Timeliness
  - The data is kept up to date.
- Consistency
  - The data agrees with itself.

Major focus of Data Cleaning process

- Dealing with
  - **Incomplete data**
  - **Noisy data**
  - **Inconsistent data**

Major focus of Data Cleaning process

- Dealing with
  - **Incomplete data**: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data…
  - **Noisy data**: containing errors, outliers…
  - **Inconsistent data**: containing discrepancies in codes or names…

Databases vs. Datasets

Measure	Databases (RDBMS)	Datasets (Data Science)
Data Value	"Precious"	"Cheap"
Data Volume	Modest	Massive
Priorities	Consistency, Error recovery, Auditability	Speed, Availability, Query richness
Structured?	Strongly (e.g. *Schema*)	Weakly or none (e.g. *Text*)
Properties	Transactions, ACID*	CAP# theorem (2/3)
Implementations / products	Dbase, Oracle, SQL	NoSQL: MongoDB, CouchDB, Hbase, Cassandra,…
Examples	Bank/Personnel/medical records, Census..	Clickstream data, GPS logs, Tweets, Building sensor readings…

*ACID = *Atomicity, Consistency, Isolation and Durability*

#CAP = *Consistency, Availability, Partition Tolerance*

ACID vs. CAP

- **Databases (ACID)**
  - Atomicity - *all transaction operations are taken as one whole unit.*
  - Consistency - *only valid data following all rules and constraints is written in the database*
  - Isolation - *transactions are securely and independently processed at the same time without interference.*
  - Durability – *committed changes remain available even in case of failure.*
- **Datasets (CAP 2/3)**
  - Consistency - *Every read receives the most recent write (or an error).*
  - Availability - *Every request receives a response (or error), without the guarantee that it contains the most recent write.*
  - Partition Tolerance - *System continues to operate despite an arbitrary number of messages being dropped (or delayed).*

Data Cleaning

Data Cleaning

Tasks -

1. Dealing with missing values
2. Identifying outliers and smoothing out noisy data
3. Correcting inconsistent data

Dealing with missing values

Dealing with Missing Values

- Types of Missing Data
  - Missing Completely At Random (MCAR)
  - Missing At Random (MAR)
  - Missing Not At Random (MNAR)

Types of Missing Data

Missing Completely At Random (MCAR)

- **No relationship** between the missing data and any variables
  - Every observation is as equally likely to be missing as any another observation.
- E.g.
  - A person accidently missed paying his parking ticket.
  - A student oversleeps and does not arrive in time to take the first section of a test.
  - Some lab test values are missing due to low battery of the equipment.

Types of Missing Data
Missing Completely At Random (MCAR)

MCAR can cause reduction in the statistical power of the data.

Most missing data treatments can be performed on MCAR data without introducing bias.

Types of Missing Data
Missing At Random (MAR)

- **No relationship** between the missing data and **independent variable** where the missingness occurs.
  - The likelihood of missingness may be related to another dataset variable.
- E.g.
  - Women report their weight less frequently than males on a survey.
  - Men are less likely to talk/answer questions about their mental well-being.
  - When placed on a soft surface, a weighing scale may produce more missing values than when placed on a hard surface.

Types of Missing Data
Missing At Random (MAR)

- **Many missing data treatments can be performed on datasets with data MAR without introducing bias.**

Types of Missing Data
Missing Not At Random (MNAR/NMAR)

- probability of an observation being missing **depends on its measured variable.**
  - the probability of being missing varies for reasons that are unknown to us.
- E.g.
  - People in lower/higher income groups are more likely not to report income on a survey.
  - Struggling readers are more likely to skip questions on a reading test.
  - In public opinion research, those with weaker opinions respond may less often.

Types of Missing Data
Missing Not At Random (MNAR/NMAR)

- the most troublesome type of missing data
- often termed "non-ignorable."

Dealing with Missing Values - Techniques

Dealing with Missing Values - Techniques

1. List-wise Deletion
2. Pair-wise Deletion
3. Fill in the missing value manually
4. Use a global constant to fill in the missing value
5. Use the most probable value to fill in the missing value
6. Imputation of Missing Data

Dealing with Missing Values - Techniques

1. List-wise Deletion

- *Process*: if any observation is missing for any data instance, delete all of the data for that instance.
- Assumes that the given data are MCAR.
- Example -

dv	iv1	iv2	iv3	iv4
80	50		NA	85
95	45	53	100	75
70	30	65	110	78
NA	42	67	105	92

⇨

dv	iv1	iv2	iv3	iv4
95	45	53	100	75
70	30	65	110	78

Dealing with Missing Values - Techniques

1. List-wise Deletion

- Pros
  - Easy to perform as the process is simple elimination.
- Cons
  - Decreases the sample size & statistical power
  - Increases standard error & widens confidence intervals

Dealing with Missing Values - Techniques

2. Pair-wise Deletion

- *Process*: remove cases that have missing data only when it pertains to a certain calculation.
  - also referred to as Available Case Analysis.
- assumes the data are MCAR.

Dealing with Missing Values - Techniques

2. Pair-wise Deletion

Example:

dv	age	weight	height
80	50	NA	58
95	45	100	62
70	30	110	NA
110	NA	105	68

⇨

dv	age	weight	height
95	45	100	62
70	30	110	NA
110	NA	105	68

* If weight is not being used in the analysis, the cases where weight is missing are not removed.

If weight is a variable in the analysis, then cases are removed.

Dealing with Missing Values - Techniques

2. Pair-wise Deletion

- Pros
  - Retains more data compared with listwise deletion
- Cons
  - Can introduce bias if data are not MCAR

Dealing with Missing Values: Techniques

3. Fill in the missing value manually
   - tedious + infeasible?
4. Use a global constant to fill in the missing value
   - What about "unknown" or "NA"?
5. Use the most probable value to fill in the missing value
   - inference-based methods such as rule based, decision tree etc.
6. Imputation of Missing Data

Imputation of Missing Data

- *Imputation* - the process of replacing missing data with substituted values based on some insights/analysis of available data.

Types:

- Single Data Imputation
- Multiple Data Imputation

Single Data Imputation Methods

- **Cold-deck Imputation**
  - Use some analysis of the variables to fill in the missing value.
    - For e.g. use measure of central tendency (mean, median, mode)
  - Values are imputed from a distribution of valid available data.
  - Can help avoid problems associated with listwise deletion of cases that have missing values.

Single Data Imputation Methods

- **Cold-deck imputation (with Data Means)**

Example

dv	iv1	iv2	iv3	iv4
80	50	NA	NA	86
95	45	54	100	76
70	30	65	110	78
NA	43	67	105	92

⇨

dv	iv1	iv2	iv3	iv4
80	50	62	105	86
95	45	54	100	76
70	30	65	110	78
82	43	67	105	92
Means:	82	42	62	105	83

Single Data Imputation Methods

- **Cold-deck Imputation**
  - Pros
    - Retains sample size
  - Cons
    - Decreases standard deviation.
    - Creates smaller confidence intervals, increasing the probability of Type 1 errors.

Single Data Imputation Methods

- **Hot-deck or Dynamic Imputation**
  - for each missing value, find an observation with similar values in column X and take its Y value.
    - Also referred to as matching.
  - If multiple matching values are found, the mean of those values is imputed.

Single Data Imputation Methods

- **Hot-deck imputation**

Example

dv	iv
90	4
NA	3
64	3.5
100	5
88	4
NA	6

dv	iv
90	4
64	3
64	3.5
100	5
88	4
100	6

Single Data Imputation Methods

- **Hot-deck Imputation**
  - Pros
    - Retains dataset size.
    - Reduces standard errors
  - Cons
    - Difficult to perform when there are multiple variables with missing data.
    - May identify more than one similar case
      - *You may have to randomly select one or use average.*

Single Data Imputation Methods

- **Last Observation Carried Forward (LOCF)**
  - imputes the last measured value of the endpoint to all subsequent, scheduled, but missing, evaluations.
    - Used in cases where data collection is interrupted before the predetermined last evaluation timepoint.

Baseline	Week 2	Week 6	Week 8	Week 10	Recorded
9	8	7	6	5	5
9	8	7	-	-	9

Single Data Imputation Methods

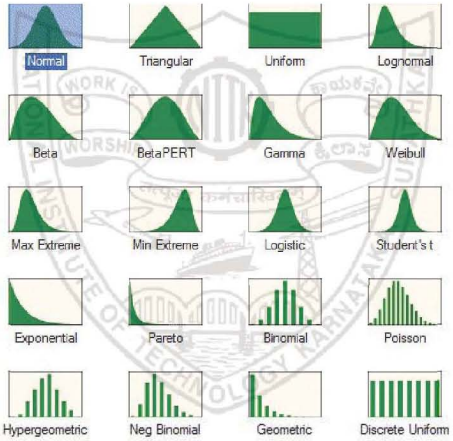
- **Next Observation Carried Backward (NOCB)**
  - Used in cases where data collection is interrupted, but later observations are available.
    - imputes the newest measured value to all earlier missed evaluations.

Baseline	Week 2	Week 6	Week 8	Week 10	Recorded
9	8	7	6	5	5
9	8	7	-	-	7

Single Data Imputation Methods

- **Distribution-based Imputation**
  - Assign value based on the probability distribution of the available data.
  - Tries to capture the "observed" empirical distribution of data.

Types of Distributions



Single Data Imputation Methods

- **Distribution-based Imputation –** *Expectation-Maximization (EM)*
  - Iterative process -
    - Make an initial guess for the model's parameters and create a probability distribution. ("E-Step" for the "Expected" distribution)
    - Feed newly observed data into the model.
    - Tweak the probability distribution from the E-step to include the new data. ("M-step")
    - Steps 2 through 4 are repeated until stability is reached.
      - *i.e. a distribution that doesn't change from the E-step to the M-step*

Single Data Imputation Methods

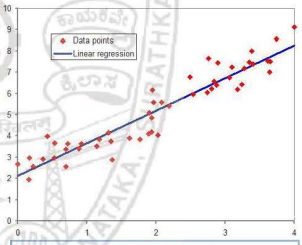
- **Distribution-based Imputation –** *Expectation-Maximization (EM)*
  - Limitations:
    - EM is very very slow, even on the fastest computer.
    - works best when missing data percentage is small and the dimensionality of the data isn't too big.
      - *higher the dimensionality, the slower the E-step;*
      - *for data with larger dimensionality, the E-step faces the problem of local maximum.*

Single Data Imputation Methods

- **Regression Imputation**
  - A technique for determining the statistical relationship between two or more variables –
    - Here, a change in a dependent variable is associated with (and depends on), a change in one or more independent variables.

Single Data Imputation Methods

- **Regression Imputation**
  - assumes that imputed values fall directly on a regression line with a nonzero slope
    - implies a correlation of 1 between the predictors and the missing outcome variable.



Regression line for 50 random points in a Gaussian distribution around the line y=1.5x+2

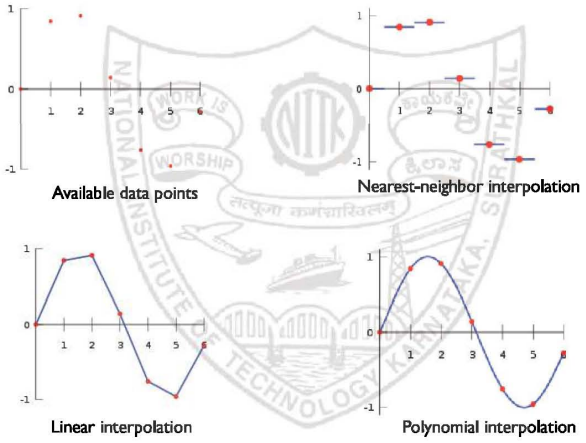
Imputation of Missing Data

- **Regression Imputation**
  - Step 1: A linear regression model is estimated on the basis of observed values in the target variable Y and some explanatory variables X.
  - Step 2: The model is used to predict values for the missing cases in Y.
  - Step 3: Missing values of Y are then replaced on the basis of these predictions.

Single Data Imputation Methods

- **Interpolation**
  - Used for constructing new data points within the range of a discrete set of known data points.

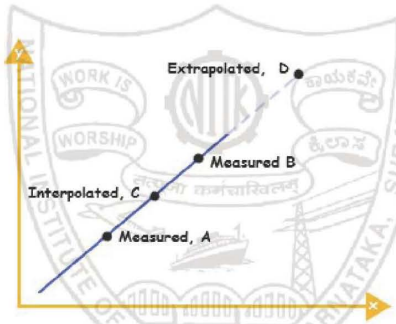
Interpolation



Single Data Imputation Methods

- **Extrapolation**
  - process of estimating, beyond the original observation range, the value of a variable on the basis of its relationship with another variable.
  - subject to greater uncertainty and a higher risk of producing meaningless results.

Extrapolation based Imputation



Multiple Data Imputation

Multiple Data Imputation

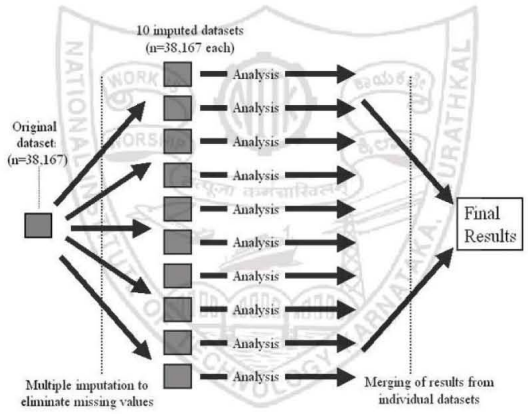
- Effective technique for dealing with MNAR datasets.
  - each missing value is replaced with multiple plausible values.
    - creates multiple potential datasets.
  - Then, these datasets are analyzed and results are compared to determine the best possible data values.

Multiple Data Imputation

- Process:
  1. Generate multiple complete-case datasets (*imputations*) through simulation (only 5 – 10 are needed)
  2. Perform analyses on each imputation
  3. Combine the multiple analyses using a set of special rules (e.g. Rubin's rules)

Rubin, Donald B. "Multiple imputation after 18+ years." *Journal of the American statistical Association* 91.434 (1996): 473-489.

Multiple Imputation - *Process*



Multiple Data Imputation

- **Step 1 : Generate multiple complete-case datasets (imputations) through simulation (about 5 – 10)**
  - Imputations are generated through maximum-likelihood based methods, e.g., *Monte Carlo simulations*
  - For each imputation, estimate missing values of each variable iteratively.
  - Continue until the means and standard deviations of the imputed values across the multiple imputed datasets start to converge.

Monte Carlo Simulations

- Provide a range of possible outcomes and the probabilities they will occur for any choice of action.
- shows the extreme possibilities —
  - the most conservative decision
  - the outcomes of worst decision.
  - ....
  - along with all possible consequences for middle-of-the-road decisions.

Monte Carlo Simulations (contd.)

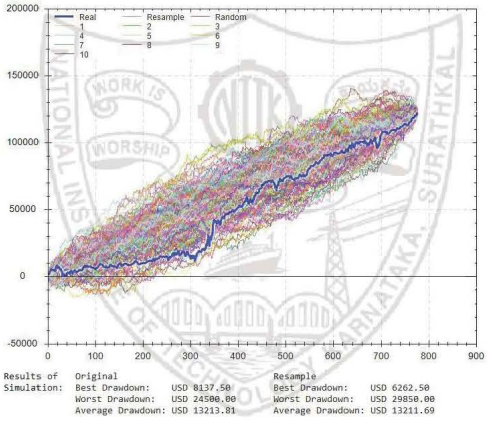
- relies on repeated random sampling to obtain numerical results.
  - use randomness to solve problems that might be deterministic in principle, but often not in practice.
  - applied to quantitative risk analysis and decision making problems.

Multiple Data Imputation *(process)*

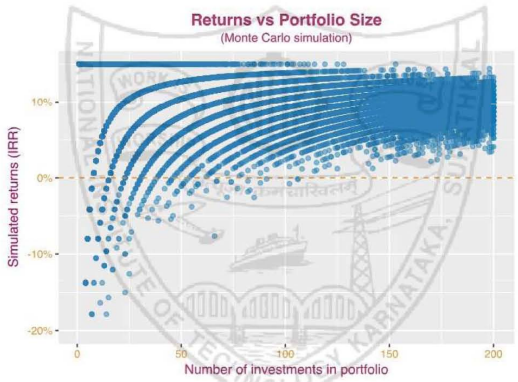
- **Step 2 : Perform planned analyses as usual on each imputed dataset**
- **Step 3: Combine the multiple analyses.**

Rubin, Donald B. "Multiple imputation after 18+ years." *Journal of the American statistical Association* 91.434 (1996): 473-489.

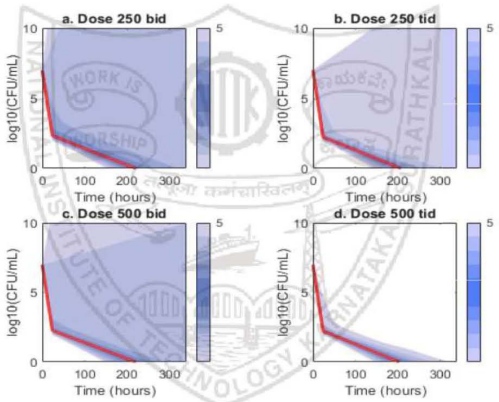
Monte Carlo Simulation – NASDAQ stock closing



Monte Carlo Simulation –Mutual Fund Management



Monte Carlo Simulation – Drug dose estimation (link)



Monte Carlo Simulation – Target area estimation

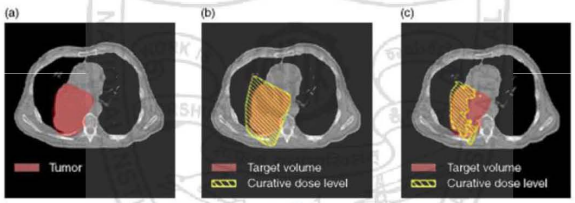
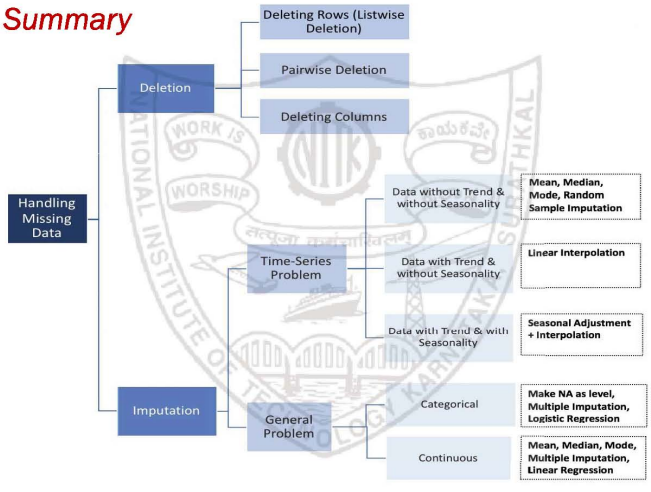


Figure (a) Axial view of a tumor in the left lung.
(b) The most accurate dose calculation available in clinics today indicates complete coverage of the tumor with a curative dose.
(c) PEREGRINE's Monte Carlo calculation of the intended dose in (b) reveals significant underdosing would occur near the boundaries of the tumor, with increased likelihood of recurrence.

Monte Carlo Simulations – Illustrative examples

- Monte Carlo Simulation for multiple problems illustrated
  - https://www.youtube.com/watch?v=7ESK5SaP-bc
  - https://www.youtube.com/watch?v=QfuRdbcZ1wE

Summary



More Reading

- Little, Roderick JA, and Donald B. Rubin. *Statistical analysis with missing data*. Vol. 793. John Wiley & Sons
- Royston, Patrick. "Multiple imputation of missing values." *The Stata Journal* 4.3 (2004): 227-241.
- Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual review of psychology*, *60*, 549-576.