Missing value and error in data [redacted]

Missing Value Handling

- Missing data reduces the representativeness of the sample.
- Makes it difficult to process the data for many analysis models / algorithms.
- Three main approaches to deal with missing values-
  1. Imputation
  2. Omission
  3. Analysis

Missing value and error in data [redacted]

Missing Value Handling

- **Imputation**
  - Values are filled in the place of missing data
  - Works well for situation where analysis tools are not robust to missing values
  - Dataset sizes are not reduced but noise gets imposed with the imputation

Estimation methods: regression, maximum likelihood estimation and approximate Bayesian bootstrap

Missing value and error in data [redacted]

Missing Value Handling

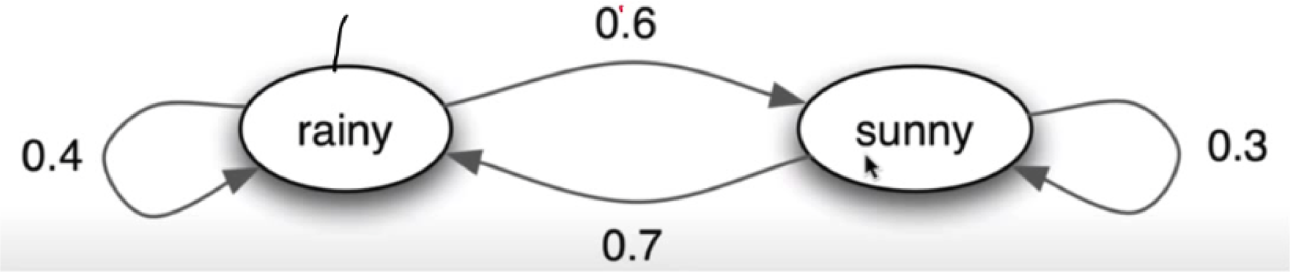
- **Omission**
  - Samples with invalid data are discarded from further analysis
  - Creates a subset of dataset with no missing values
  - Works well for models that are not robust against data missingness

Example techniques: list-wise deletion, pair-wise deletion

Missing value and error in data [redacted]

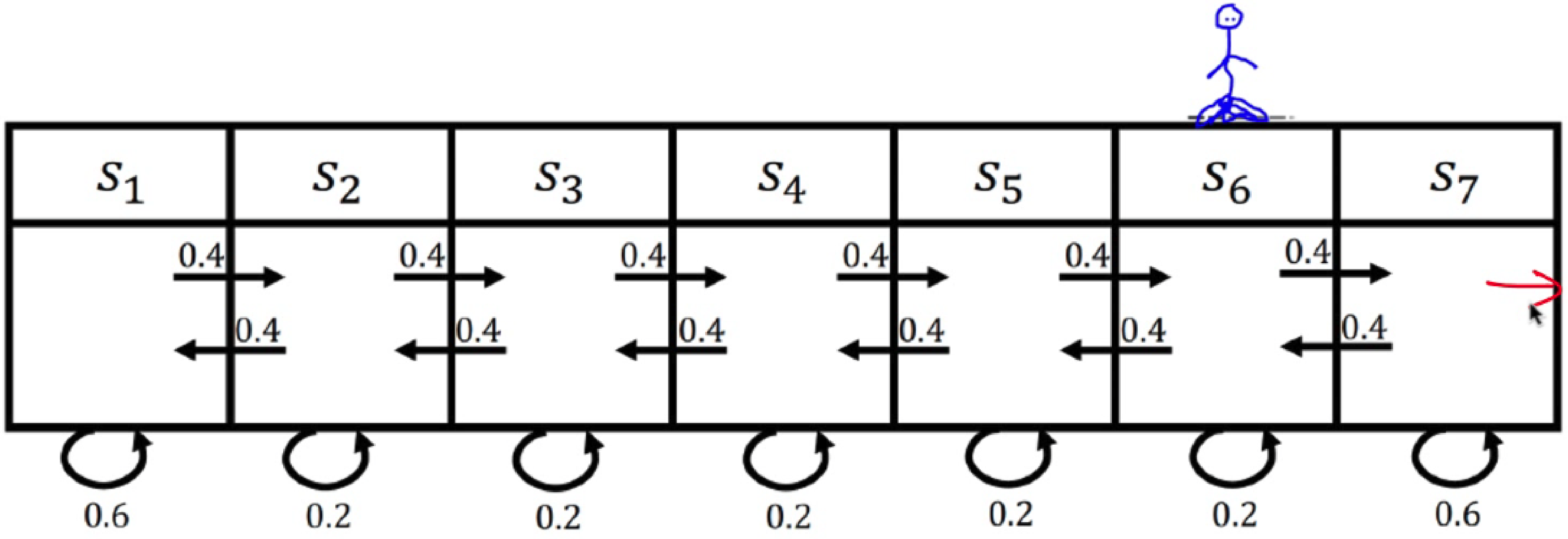
Missing Value Handling

- **Analysis**
  - Samples with invalid data are discarded from further analysis
  - Model-based techniques used to determine missing values
  - Various non-stationary Markov chain models can be used for time series data



Missing value and error in data [redacted]

Analysis: Markov chain models can be used for time series data



Missing value and error in data [redacted]

Error in Data

- The difference between the recorded data and true value
- Higher error rates in data makes it less representative

Types

- The major types of data error includes-

- Sampling error
- Non-sampling error

Missing value and error in data

[redacted]

Sampling error

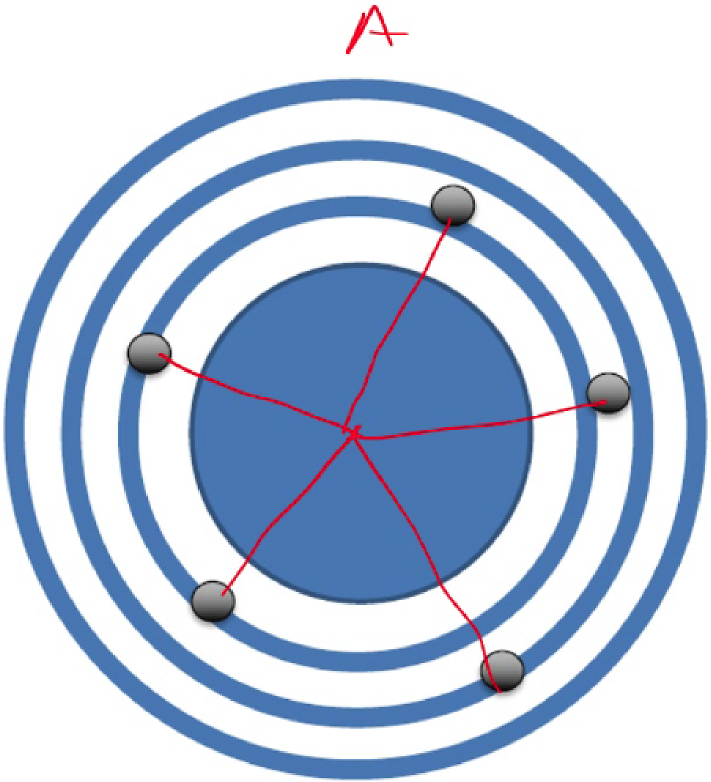
- Error for using data from sample of the population, in place of entire population
- It's the difference between the estimate from the sample and true vale for the population
- Could occur for very small sample size
- If the sampling is not random and have some bias

Theory of Measurements

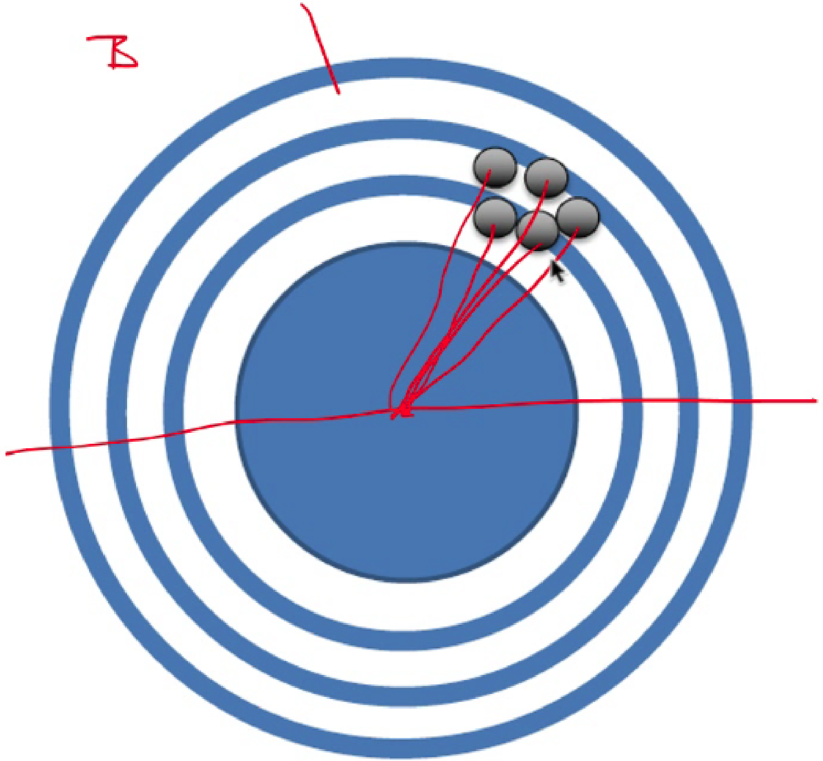
[redacted]

Accuracy & Precision

Low Accuracy, Low Precision



Low Accuracy, High Precision



Relational Database

[redacted]

Relational Model

- Relational model uses table (called relation) to represent a collection of related data values
- Rows are called records or tuples
- Columns are called attributes
- The number of attributes (i.e., number of columns) is called the degree
- Example database: MySQL, PostgreSQL and SQLite3

[redacted]

Non-relational Database

Non-relational database

- Database that does not use the tabular schema of rows and columns; i.e. it don’t use relational model.
- Often refers to NOSQL (not only SQL); Data may be stored as
  - simple key/value pairs
  - JSON documents or
  - a graph consisting of edges and vertices.
- Most NOSQL systems are distributed databases or distributed storage systems
- Example DB: MongoDB, Oracle NoSQL, Apache CouchDB and Redis.

[redacted]

Non-relational Database

NOSQL Systems

- Database NOSQL systems focus on storage of “big data”
- Typical applications that use NOSQL
    - Social media
    - Web links
    - Marketing and sales
    - Posts and tweets
    - Road maps and spatial data
    - Email