Quantitative Data Analysis: Terminologies

Kyungsik Han

Terminology

- Data type
  - Continuous
    - Data that can take any value within a certain range
  - Discrete
    - Data that can only take integer values such as number of times
  - Categorical
    - Data that can only take values within possible categories
  - Binary
    - Special case of categorical data with only two values
  - Ordinal
    - Categorical data with a clear order between values

Terminology

- Data frame
    - Data frame
        - The tabular data structure that is the most basic in statistics and machine learning models
    - Feature
        - Each column in the table represents one feature
    - Outcome
        - Use features to predict results in experiments or research
    - Record
        - Each row of the table represents one record

Terminology for Stats

- Metrics and estimated values
  - Statisticians usually use the term **estimate** to refer to values computed from data
  - This is to distinguish values from theoretical true values which represent the actual state
  - On the other hand, data scientists or business analysts call them metrics

Terminology for Stats

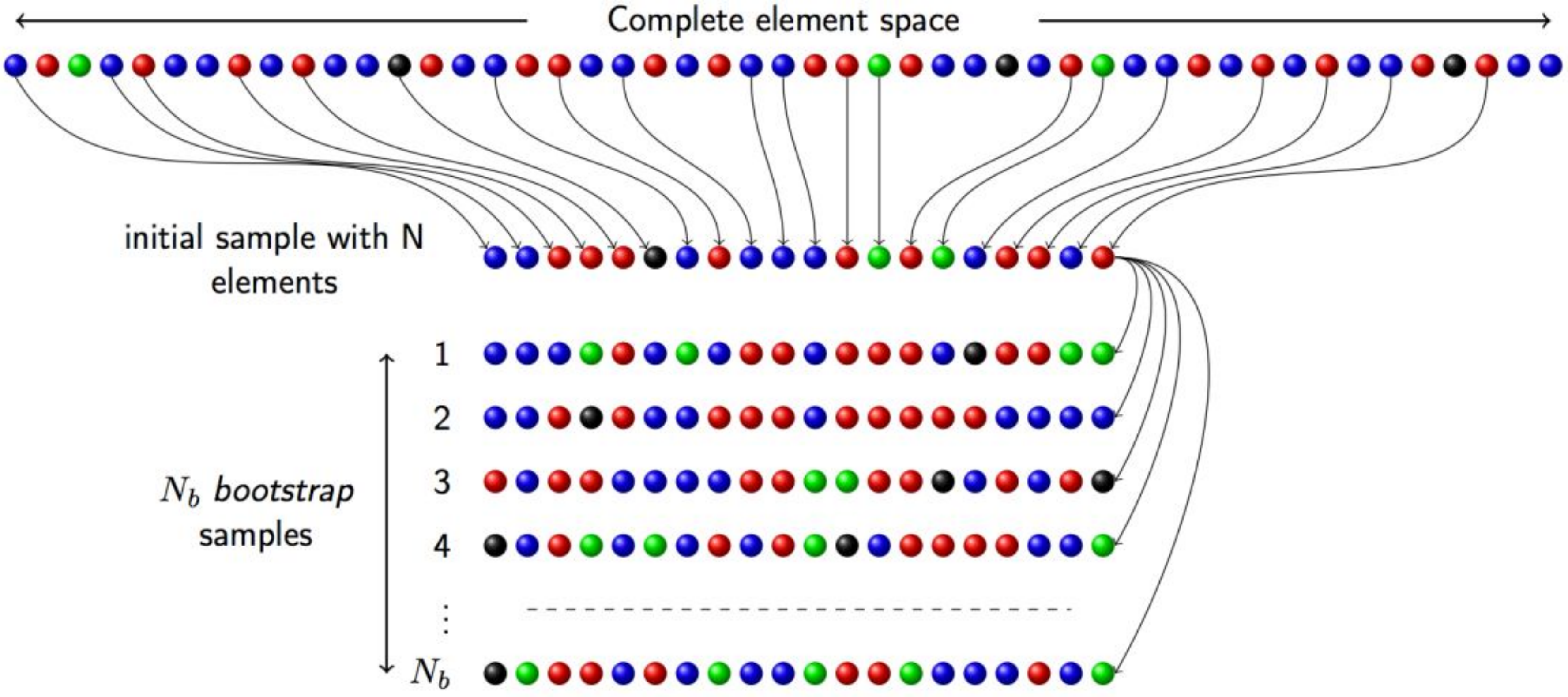
- Sampling
    - Sample
        - a subset obtained from a larger dataset
    - Population
        - an entire target or whole set of datasets
    - N(n)
        - the size of the population (sample)
    - Random sampling
    - Stratified sampling
        - divide the population into layers, then randomly sample from each layer
    - Simple random sample
        - A sample obtained by random sampling without population stratification
    - Sample bias
        - A sample that incorrectly represents a population

Terminology for Stats

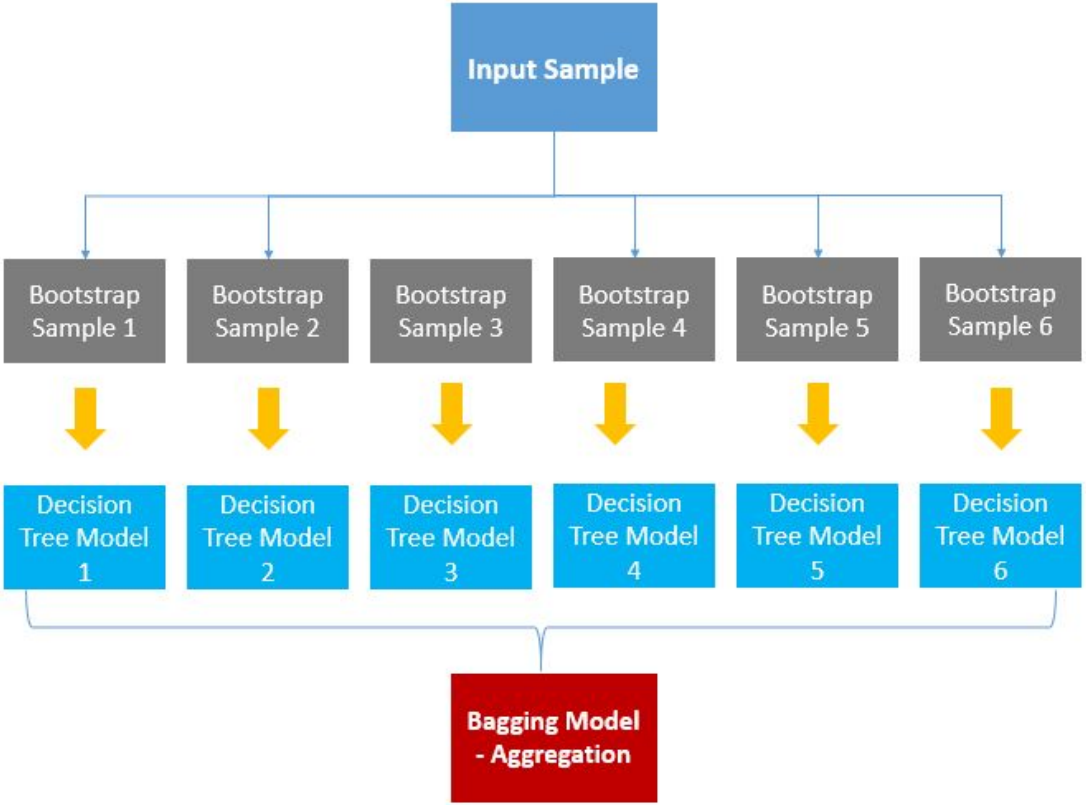
- Sample distribution
  - Sample statistics
    - metrics obtained from sample data from a larger population
  - Data distribution
    - the frequency distribution of each individual value in a data set
  - Sampling distribution
    - the frequency distribution of sample statistics from multiple samples or resamples
  - Central limit theorem
    - As the sample size increases, the sample distribution tends to follow the normal distribution
  - Standard error
    - the variance of a sample statistic from multiple samples (not to be confused with the standard deviation, which means the variance of individual data values)

Terminology for Stats

- Bootstrap
    - Bootstrap sample
        - a recovery sampling sample obtained from a set of observed data
        - Replicating the original sample thousands or millions of times, which results in a virtual population that contains all the information from the original sample
        - Samples can be collected for the purpose of estimating the distribution of samples from this hypothetical population
        - Used in decision-making trees
            - This process is called bagging
    - Re-sampling
        - The process of repeatedly sampling from observational data



bootstrapping



Input Sample

Bootstrap Sample 1

Bootstrap Sample 2

Bootstrap Sample 3

Bootstrap Sample 4

Bootstrap Sample 5

Bootstrap Sample 6

Decision Tree Model 1

Decision Tree Model 2

Decision Tree Model 3

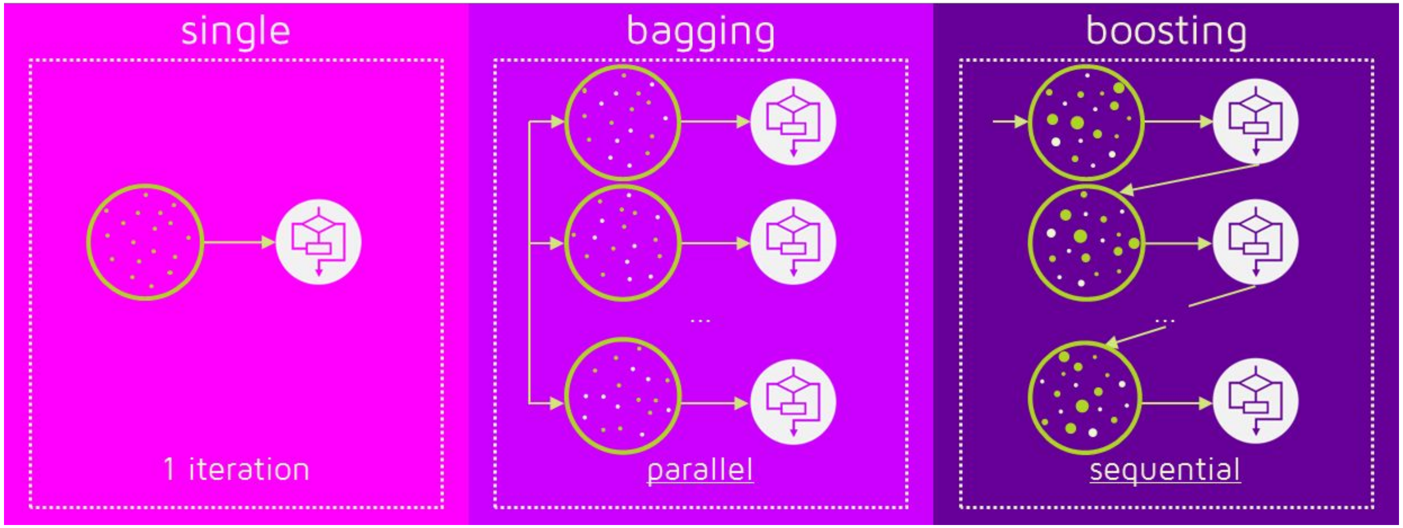
Decision Tree Model 4

Decision Tree Model 5

Decision Tree Model 6

Bagging Model - Aggregation

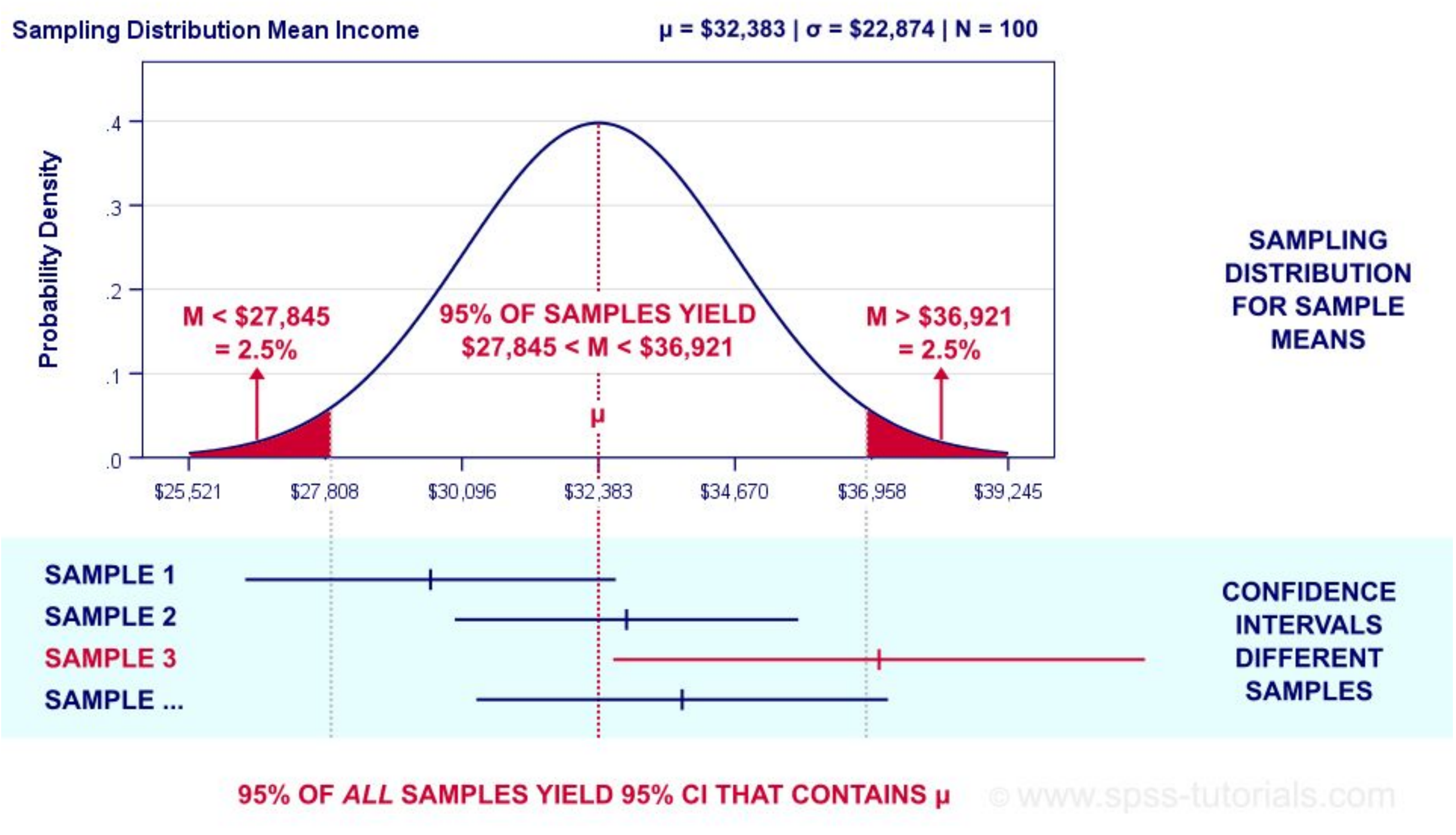
bagging vs boosting



Terminology for Stats

- Confidence interval
  - Confidence level
    - Percentage of confidence intervals expected to contain statistics of interest, obtained in the same way from the same population
  - Interval endpoint
    - the highest and lowest confidence intervals

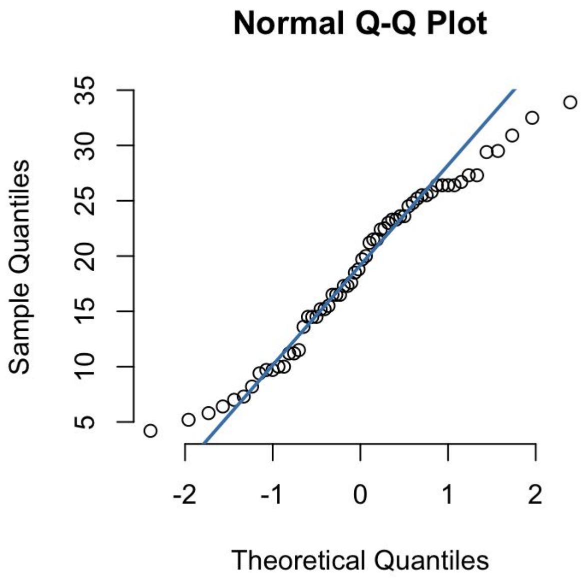
Terminology for Stats



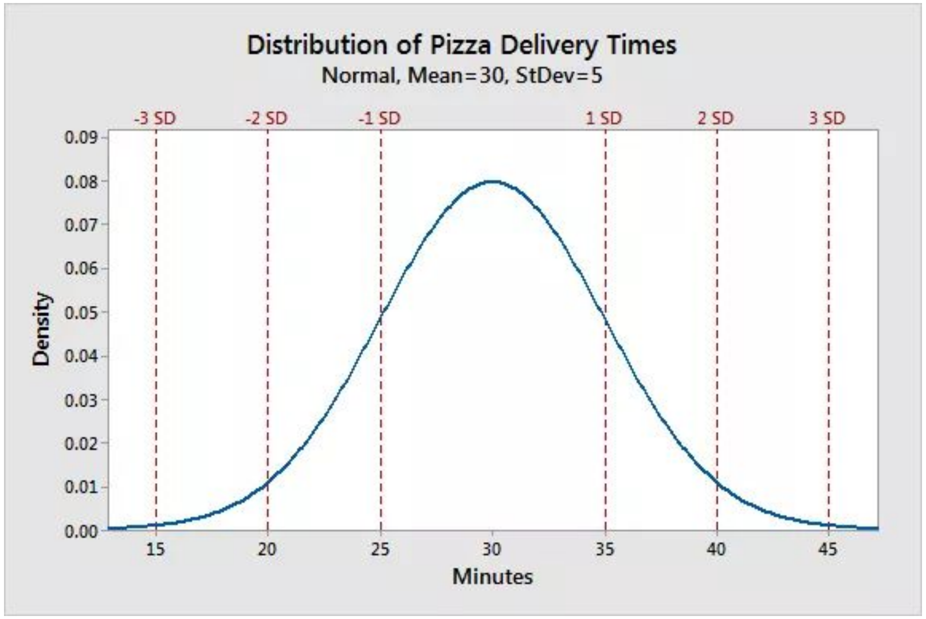
Terminology for Stats

- Normal distribution
  - Error
    - the difference between a data point and a predicted value or average
  - Standardize
    - subtract the mean and divide by the standard deviation
  - z-score
    - the result of normalizing individual data points
  - Standard normal distribution
    - normal distribution with mean = 0, standard deviation = 1
  - QQ-plot
    - a plot showing how close the sample distribution is to the normal distribution

Terminology for Stats



QQ Plot



Normal distribution

Controlled Experiment Terminology

Controlled Experiment Terminology

- Factor
  - An independent variable (e.g., input device)
- Levels
  - Possible values of a factor (e.g., *touchpad* and *trackball* are two levels of the factor input device)
- Between-subjects factor
  - A factor for which each subject performs with one level (e.g., each subject uses the *touchpad* or the *trackball* but not both)
- Within-subjects factor
  - A factor for which each subject performs with all levels (e.g., each subject uses the *touchpad* and the *trackball*)

Controlled Experiment Terminology

- Counterbalance
  - Ordering the levels of a factor so as to avoid confounding the results (e.g., making sure half of the subjects do *touchpad* first, and half do *trackball* first in a within-subjects design)
- ANOVA
  - Abbreviation for “analysis of variance,” which is a common statistical method used to determine if there are differences between levels of different factors (more than two levels)
- t-test
  - A simple statistical test to compare the means and distributions of two groups (of two levels of a single factor) (e.g., *touchpad* vs *trackball* throughput)
- p-value
  - The result of a statistical test. By convention, a p-value less than 0.05 is deemed “statistically significant”