Data preparation and preprocessing

Chapter 2: Data Preprocessing
(book slide)

- Why preprocess the data?
- Descriptive data summarization
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation

Why Data Preprocessing? (book slide)

- Data in the real world is dirty

– **incomplete:** lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  - e.g., occupation=“ ”

– **noisy:** containing errors or outliers
  - e.g., Salary=“-10”

– **inconsistent:** containing discrepancies in codes or names
  - e.g., Age=“42” Birthday=“03/07/1997”
  - e.g., Was rating “1,2,3”, now rating “A, B, C”

Source of Dirty Data

- **Incomplete data may**
  - "Not applicable" data value when collected
  - Different considerations between the time when the data was collected and when it is analyzed.
  - Human/hardware/software problems
- **Noisy data (incorrect values) may come from**
  - Faulty data collection instruments
  - Human or computer error at data entry
  - Errors in data transmission
- **Inconsistent data may come from**
  - Different data sources
  - Functional dependency violation (e.g., modify some linked data)
- **Duplicate records also need data cleaning**

TYPES OF DATA

- Generally we distinguish:

Quantitative Data Qualitative Data

- Bivaluated: often very useful
- Remember: Null Values are not applicable
- Missing data usually not acceptable

Why Prepare Data?

- Some data preparation is needed for all mining tools
- The purpose of preparation is to transform data sets so that their information content is best exposed to the mining tool
- Error prediction rate should be lower (or the same) after the preparation as before it

Why Prepare Data?

- Preparing data also prepares the miner so that when using prepared data the miner produces better models, faster

- GIGO - good data is a prerequisite for producing effective models of any type

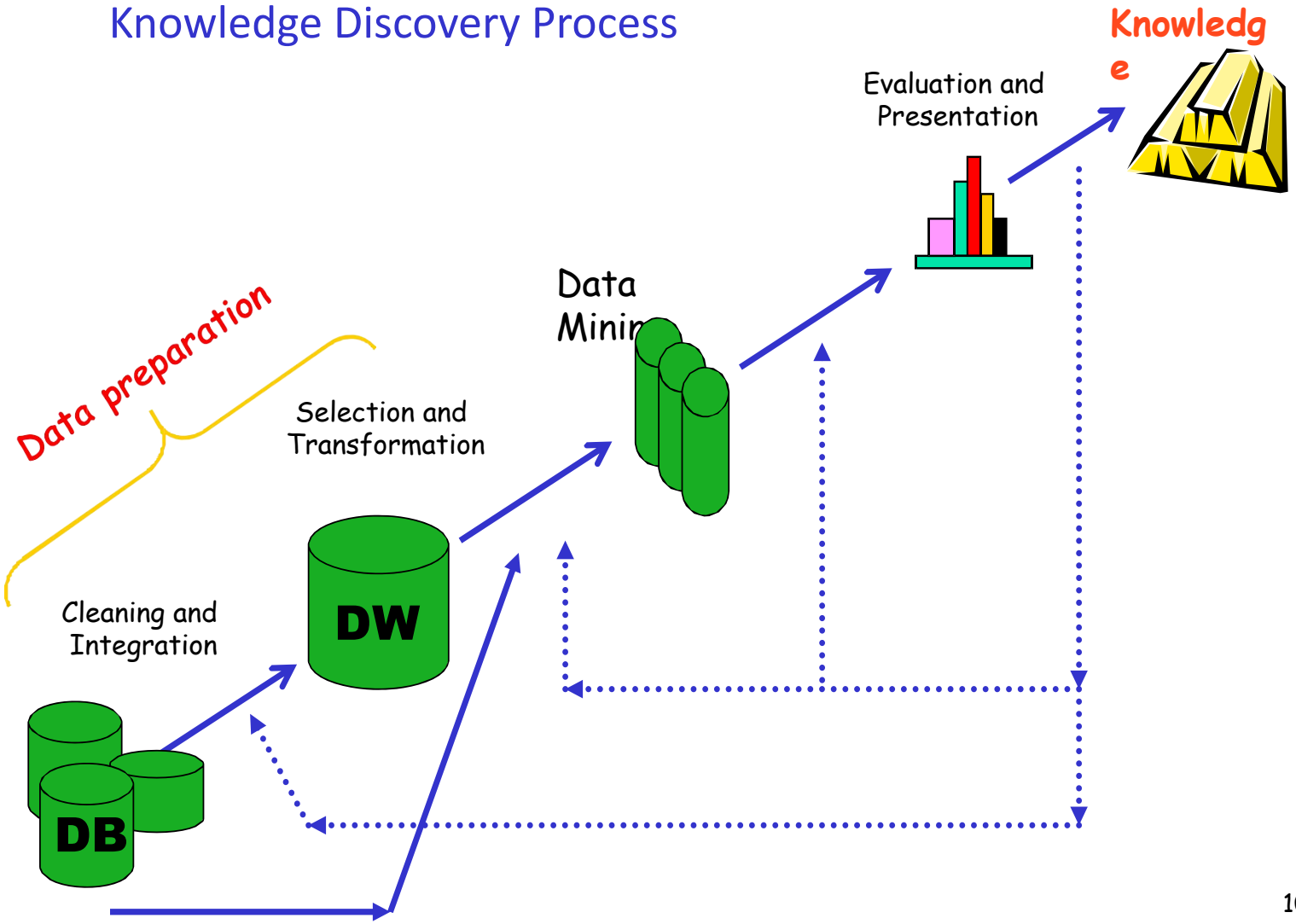
Why Prepare Data?

- Data need to be formatted for a given software tool
- Data need to be made adequate for a given method
- Data in the real world is dirty
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., occupation=""
  - noisy: containing errors or outliers
    - e.g., Salary="-10", Age="222"
  - inconsistent: containing discrepancies in codes or names
    - e.g., Age="42" Birthday="03/07/1997"
    - e.g., Was rating "1,2,3", now rating "A, B, C"
    - e.g., discrepancy between duplicate records
    - e.g., ***Endereço:*** travessa da Igreja de Nevogilde ***Freguesia:*** Paranhos

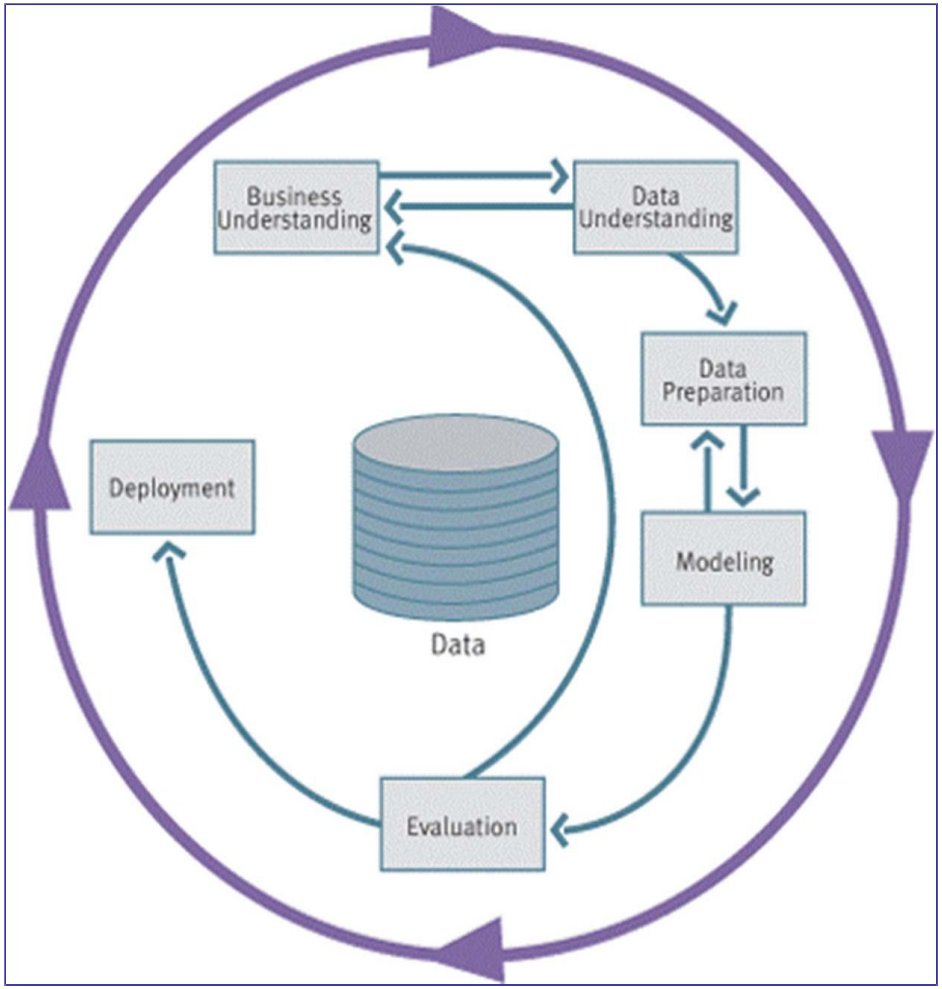
Measures of Data Quality

- A well-accepted multidimensional view of data quality:
  - Accuracy
  - Completeness
  - Consistency
  - Timeliness
  - Believability
  - Interpretability
  - Accessibility

Data Preparation as a step in the Knowledge Discovery Process

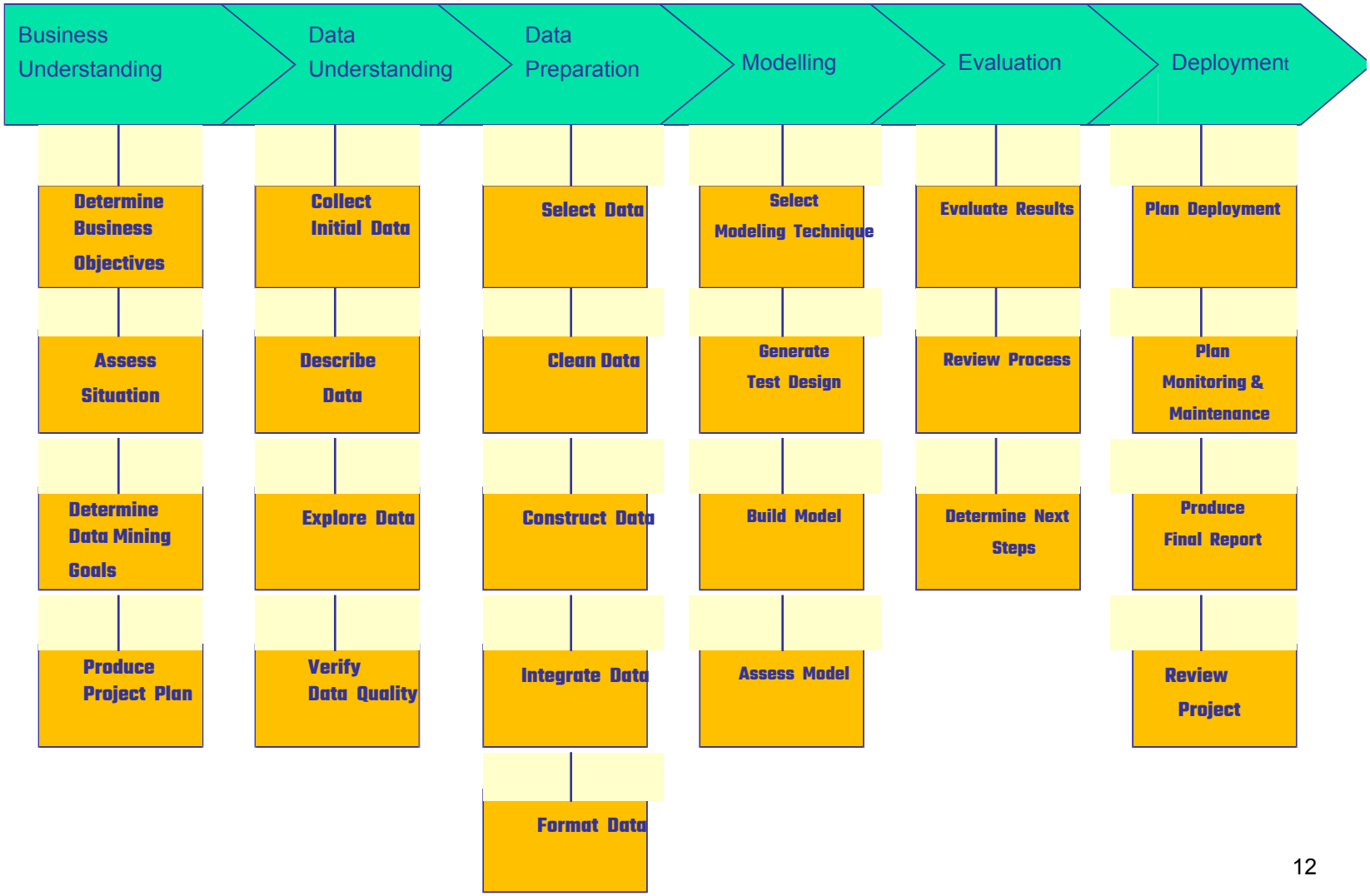


CRISP-DM

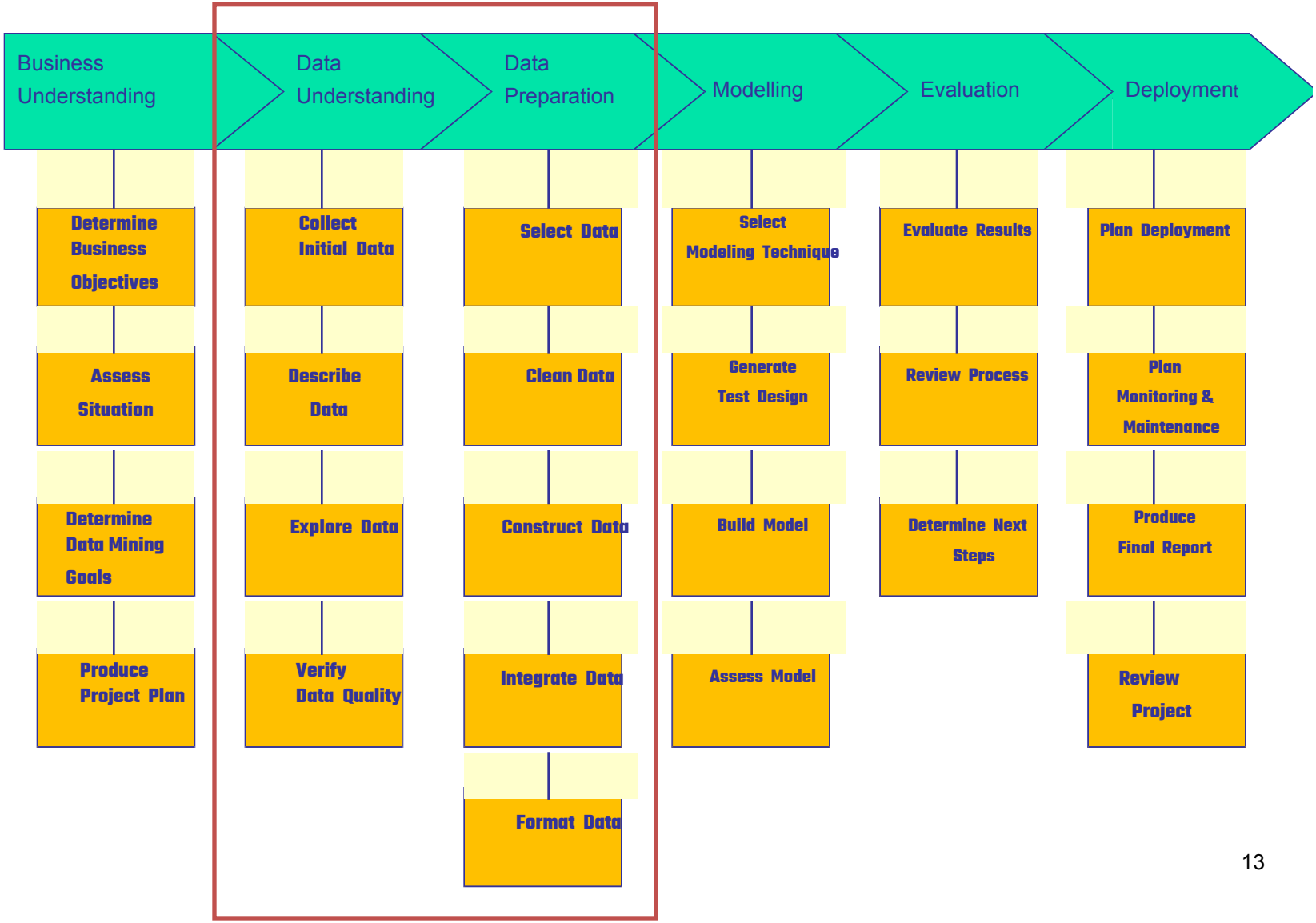


- A comprehensive data mining methodology and process model that provides anyone—from novices to data mining experts—with a complete blueprint for conducting a data mining project.
- A methodology enumerates the steps to reproduce success

CRISP-DM Phases and Tasks



CRISP-DM Phases and Tasks



CRISP-DM: Data Understanding

- **Collect data**
  - List the datasets acquired (locations, methods used to acquire, problems encountered and solutions achieved).
- **Describe data**
  - Check data volume and examine its gross properties.
  - Accessibility and availability of attributes. Attribute types, range, correlations, the identities.
  - Understand the meaning of each attribute and attribute value in business terms.
  - For each attribute, compute basic statistics (e.g., distribution, average, max, min, standard deviation, variance, mode, skewness).

CRISP-DM: Data Understanding

- **Explore data**
  - Analyze properties of interesting attributes in detail.
    - *Distribution, relations between pairs or small numbers of attributes, properties of significant sub-populations, simple statistical analyses.*
- **Verify data quality**
  - Identify special values and catalogue their meaning.
  - Does it cover all the cases required? Does it contain errors and how common are they?
  - Identify missing attributes and blank fields. Meaning of missing data.
  - Do the meanings of attributes and contained values fit together?
  - Check spelling of values *(e.g., same value but sometime beginning with a lower case letter, sometimes with an upper case letter).*
  - Check for plausibility of values, e.g. all fields have the same or nearly the same values.

CRISP-DM: Data Preparation

- **Select data**
  - Reconsider data selection criteria.
  - Decide which dataset will be used.
  - Collect appropriate additional data (internal or external).
  - Consider use of sampling techniques.
  - Explain why certain data was included or excluded.
- **Clean data**
  - Correct, remove or ignore noise.
  - Decide how to deal with special values and their meaning *(99 for marital status).*
  - Aggregation level, missing values, etc.
  - Outliers?

CRISP-DM: Data Preparation

- **Construct data**
  - Derived attributes.
  - Background knowledge.
  - How can missing attributes be constructed or imputed?
  - **Data transformation**
    - Normalization and aggregation
  - **Data reduction**
    - Obtains reduced representation in volume but produces the same or similar analytical results
  - **Data discretization**
    - Part of data reduction but reduces the number of values of the attributes;
    - particular importance especially for numerical data

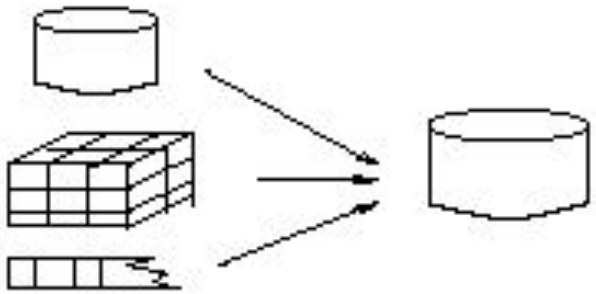
Data Transformation -2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

Data Reduction

CRISP-DM: Data Preparation

- **Construct data**
  - Derived attributes.
  - Background knowledge.
  - How can missing attributes be constructed or imputed?
- **Integrate data**
  - Integrate sources and store result (new tables and records).

Data Integration



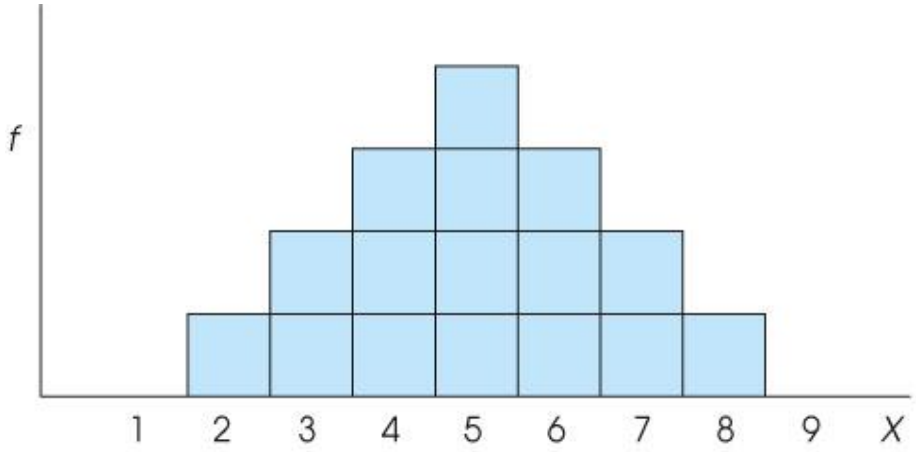
CRISP-DM: Data Preparation

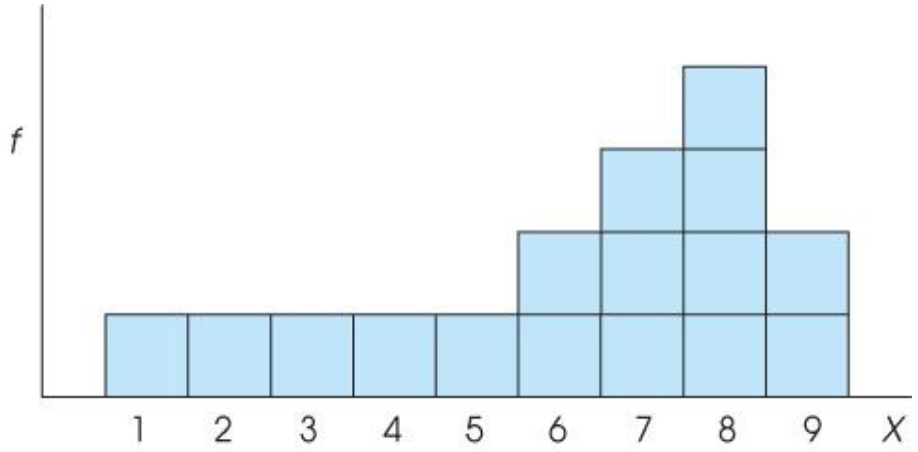
- **Format Data**
  - Rearranging attributes *(Some tools have requirements on the order of the attributes, e.g. first field being a unique identifier for each record or last field being the outcome field the model is to predict).*
  - Reordering records *(Perhaps the modelling tool requires that the records be sorted according to the value of the outcome attribute).*
  - Reformatted within-value *(These are purely syntactic changes made to satisfy the requirements of the specific modelling tool, remove illegal characters, upper case, lowercase)*

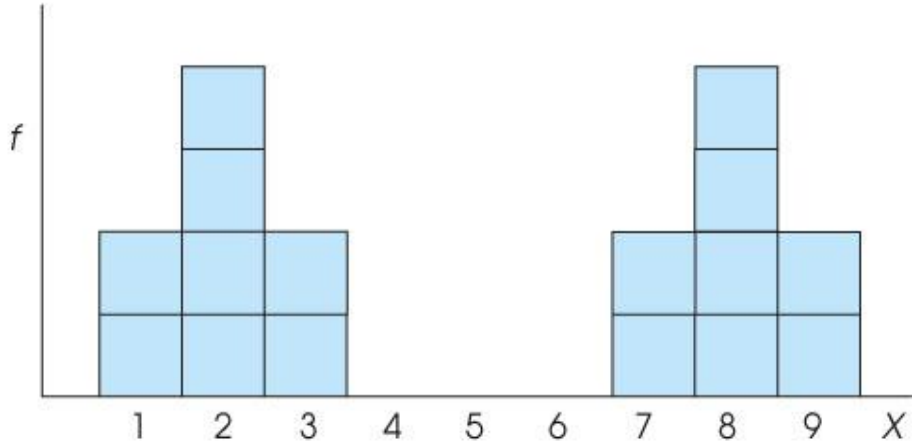
Studying Data Quality and characteristics

Descriptive Data Summarization

- Descriptive Data Summarization techniques
  - used to identify the typical properties of the data and highlight which data values should be treated as niose or outliers
- Impotent in Learning the CENTRALTENDENCY AND DISPERSION OF THE DATA
- Central Tendency
  - a statistical measure that determines a single value that accurately describes the center of the distribution and represents the entire distribution of scores
  - Goal is to identify the single value that is the best representative for the entire set of data
  - allows researchers to summarize or condense a large set of data into a single value and describe or present a set of data in a very simplified, concise form.
  - to compare two (or more) sets of data by simply comparing the average score (central tendency) of each sets


f
1
2
3
4
5
6
7
8
9
X


f
1
2
3
4
5
6
7
8
9
X


f
1
2
3
4
5
6
7
8
9
X

Descriptive Data Summarization

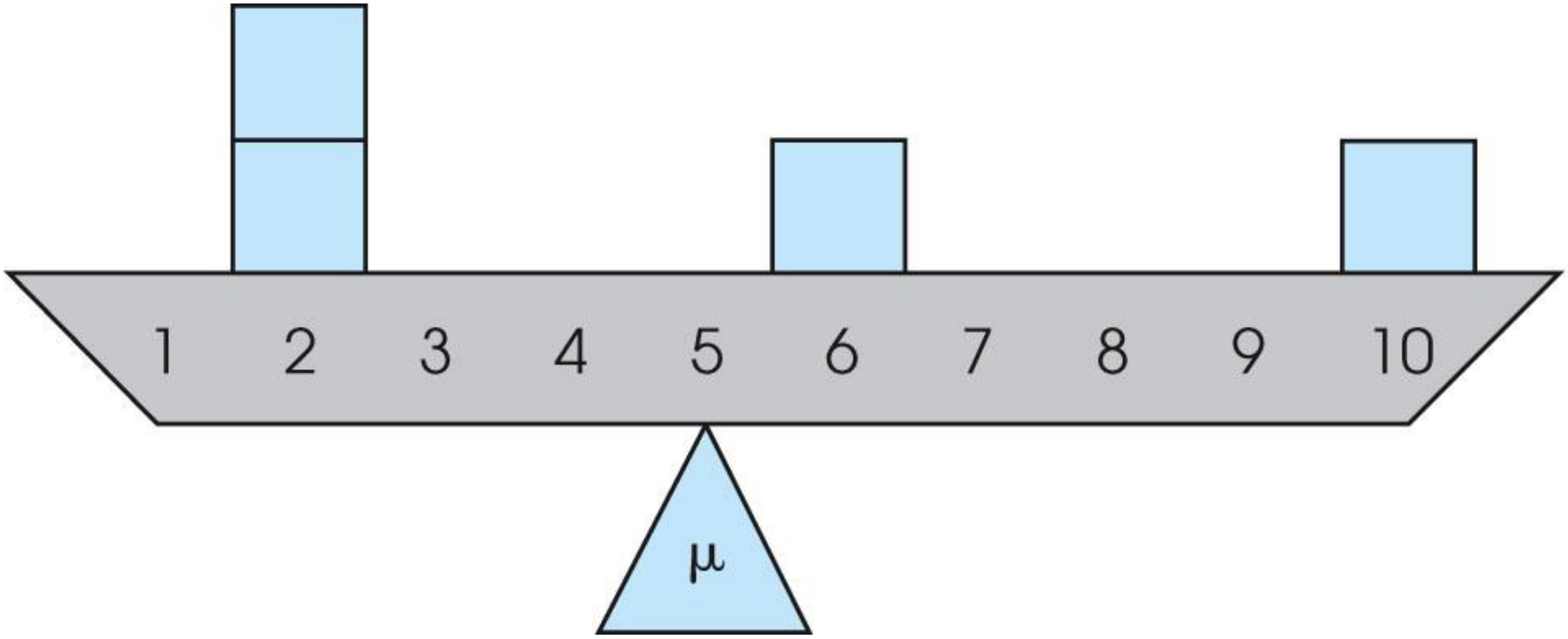
- Measures of CENTRALTENDENCY
  - Researchers use multiple statistics
  - mean, median, mode, and midrange.
- Dispersion, or variance of the data is the degree to which numerical data tend to spread.
- Measuresof Data Dispersion :
  - Range, the five-number summary (based on quartiles), the interquartile range, and standard deviation.

The Mean : Central Tendency measure

- The most commonly used measure of central tendency.
- The numerical valued attribute and values measured on an **interval or ratio scale**
- **MEAN is a distributive measure**,
  - it can be computed on subsets , and results merged in one.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \mu = \frac{\sum x}{N}$$

- MEAN is also an algebraic measure
  - can be computed by applying an algebraic function to one or more distributive measures
  - **The mean is the balance point of the distribution because the sum of the distances below the mean is exactly equal to the sum of the distances above the mean**


1
2
3
4
5
6
7
8
9
10
μ

Measuring the Central Tendency

- Sometimes each value xi may be associated with a weight wi;
- the weights reflect the significance, importance, or occurrence frequency attached to their respective values;
- In this case we compute the
- Weighted arithmetic mean, or weighted average:

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

Changing the Mean

- The calculation of the mean involves every score in the distribution
  - changing the value of any score will change the value of the mean.
- Modifying a distribution by discarding scores or by adding new scores will usually change the value of the mean.
- To determine how the mean will be affected for any specific situation you must consider:
  - 1) how the number of scores is affected
  - 2) how the sum of the scores is affected
    - If value is added to every score in a distribution, then the same constant value is added to the mean
    - if every score is multiplied by a constant value, then the mean is also multiplied by the same constant value

When the Mean Won't Work

- When a distribution contains a few extreme scores (or is very skewed), the mean will be pulled toward the extremes (displaced toward the tail)
- 1,2,3,4,5,6
- 1, 2,3,4,5,100

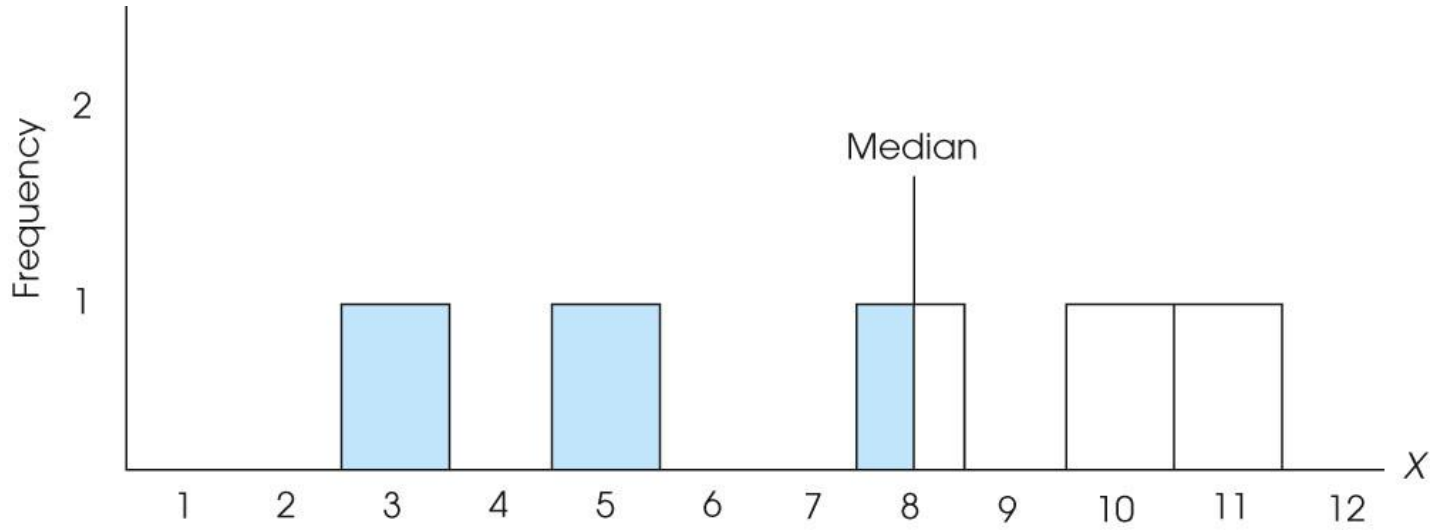
The Median

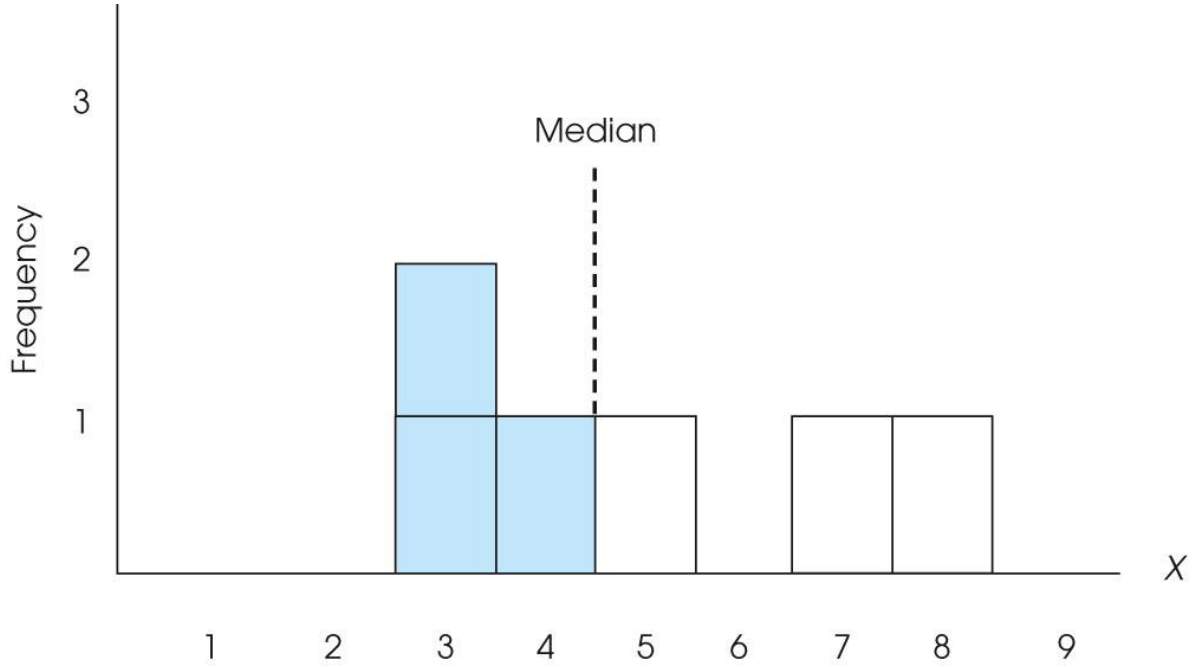
- If the **scores in a distribution are listed in order** from smallest to largest, the median is defined as the **midpoint of the list**.
- The median divides the scores so that 50% of the scores in the distribution have values that are equal to or less than the median.
- Computation of the median requires scores that can be placed in rank order (smallest to largest) and are measured on an ordinal, interval, or ratio scale.
- 1,2,3,4,5,6
- 1,2,3,4,5,100

The Median (cont.)

Usually, the median can be found by a simple counting procedure:

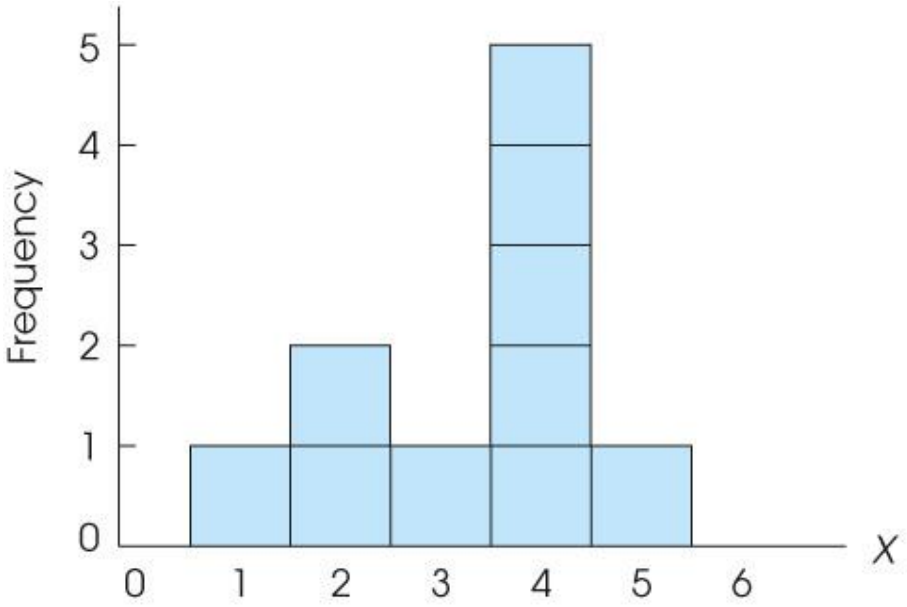
1. With an odd number of scores, list the values in order, and the median is the middle score in the list.
2. With an even number of scores, list the values in order, and the median is half-way between the middle two scores.

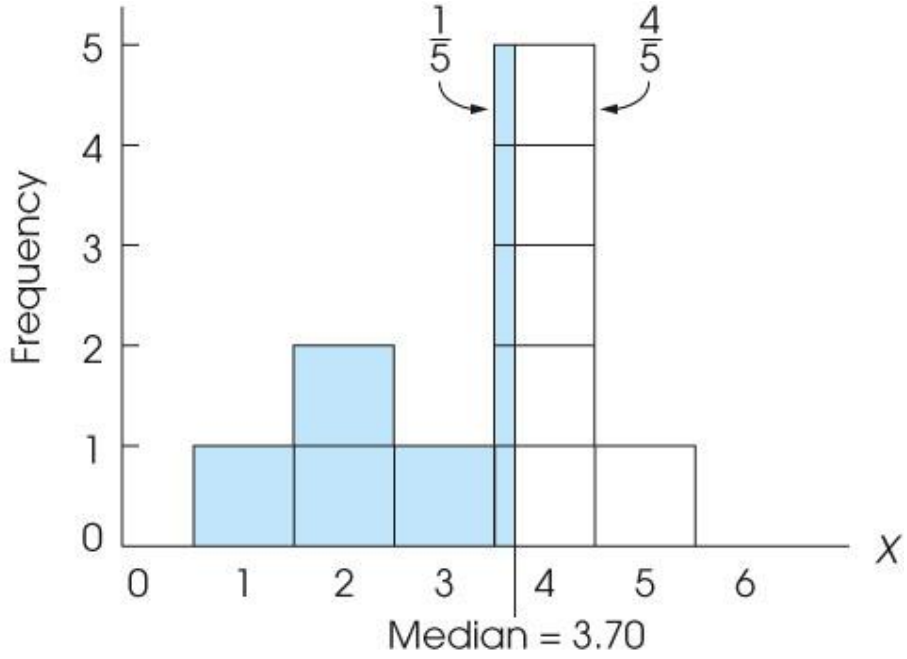

Frequency
2
1
Median
X
1
2
3
4
5
6
7
8
9
10
11
12


Frequency
3
2
1
Median
X
1
2
3
4
5
6
7
8
9

The Median (cont.)

- If the scores are measurements of a continuous variable, it is possible to find the median by **first placing the scores in a frequency distribution histogram** with each score represented by a box in the graph.
- Then, draw a vertical line through the distribution so that exactly half the boxes are on each side of the line. The median is defined by the location of the line.


Frequency
5
4
3
2
1
0
0
1
2
3
4
5
6
X


Frequency
5
4
3
2
1
0
1/5
4/5
0
1
2
3
4
5
6
X
Median = 3.70

The Median

- One advantage of the median is that it is relatively unaffected by extreme scores.
- Thus, the median tends to stay in the "center" of the distribution even **when there are a few extreme scores or when the distribution is very skewed.** In these situations, the median serves as a good alternative to the mean.

The Mode

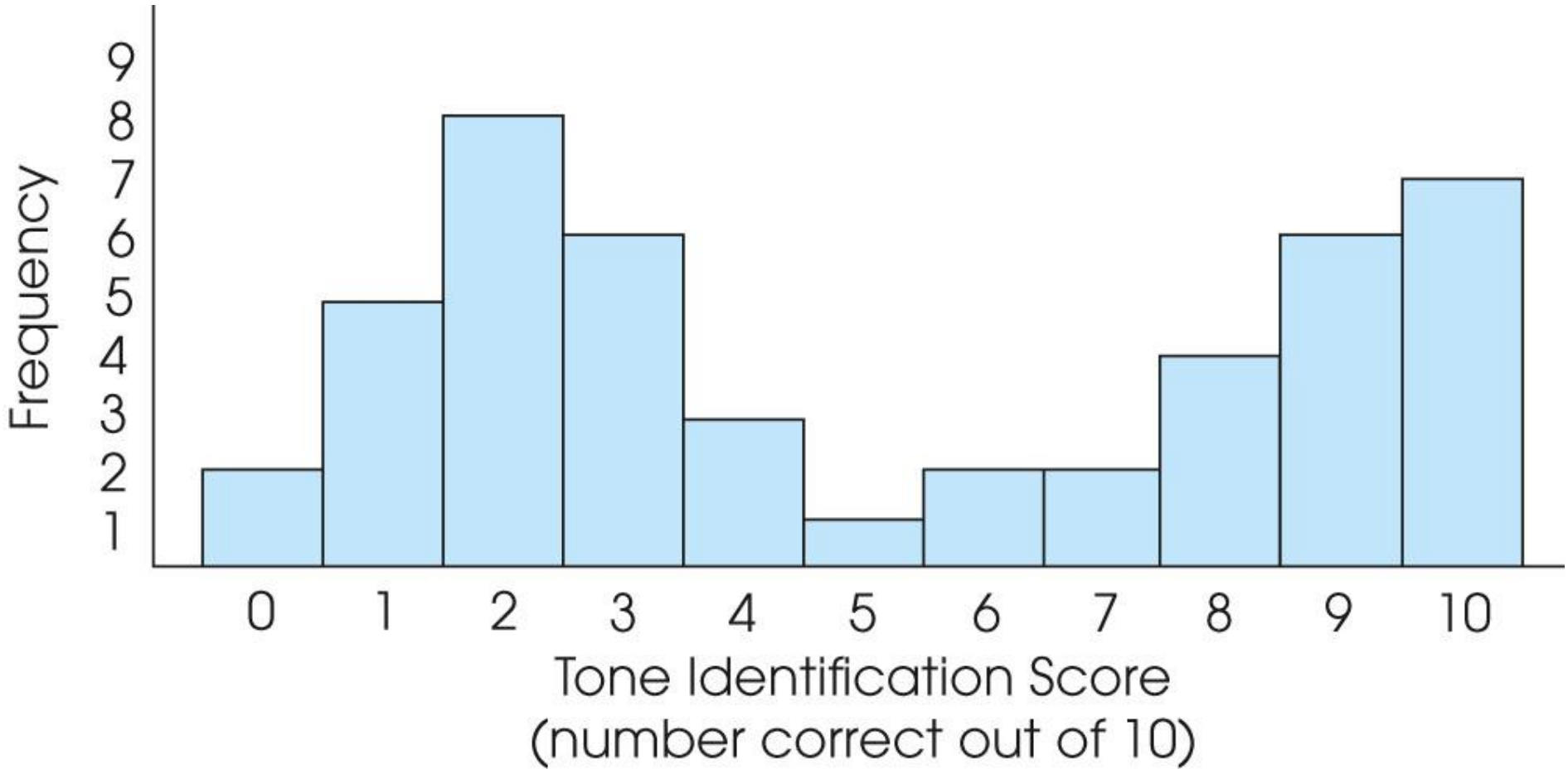
- The mode is defined as the most frequently occurring category or score in the distribution.
- In a frequency distribution graph, the mode is the category or score corresponding to the peak or high point of the distribution.
- **The mode can be determined for data measured on any scale of measurement: nominal, ordinal, interval, or ratio.**
- **0:1, 1:200, 2:300, 3:500, 4:800, 5:100, 6:1**

The Mode (cont.)

- The primary value of the mode is that it is the only measure of central tendency that can be used for data measured on a nominal scale. In addition, the mode often is used as a supplemental measure of central tendency that is reported along with the mean or the median.

Bimodal Distributions

- It is possible for a distribution to have more than one mode. Such a distribution is called **bimodal**. (Note that a distribution can have only one mean and only one median.)
- In addition, the term "mode" is often used to describe a peak in a distribution that is not really the highest point. Thus, a distribution may have a *major mode* at the highest peak and a *minor mode* at a secondary peak in a different location.


Frequency
9
8
7
6
5
4
3
2
1
0
1
2
3
4
5
6
7
8
9
10
Tone Identification Score
(number correct out of 10)

Central Tendency and the Shape of the Distribution

- Because the mean, the median, and the mode are all measuring central tendency, the three measures are often systematically related to each other.
- In a symmetrical distribution, for example, the mean and median will always be equal.

Central Tendency and the Shape of the Distribution (cont.)

- If a symmetrical distribution has only one mode, the mode, mean, and median will all have the same value.
- In a skewed distribution, the mode will be located at the peak on one side and the mean usually will be displaced toward the tail on the other side.
- The median is usually located between the mean and the mode.

Reporting Central Tendency in Research Reports

- In manuscripts and in published research reports, the sample mean is identified with the letter M.
- There is no standardized notation for reporting the median or the mode.
- In research situations where several means are obtained for different groups or for different treatment conditions, it is common to present all of the means in a single graph.

Reporting Central Tendency in Research Reports (cont.)

- The different groups or treatment conditions are listed along the horizontal axis and the means are displayed by a bar or a point above each of the groups.
- The height of the bar (or point) indicates the value of the mean for each group. Similar graphs are also used to show several medians in one display.

Measuring the Dispersion of Data (book slide)

- Dispersion, or variance of the data is the degree to which numerical data tend to spread
  - Range, the five-number summary (based on quartiles), the interquartile range, and standard deviation.

Variance and SD

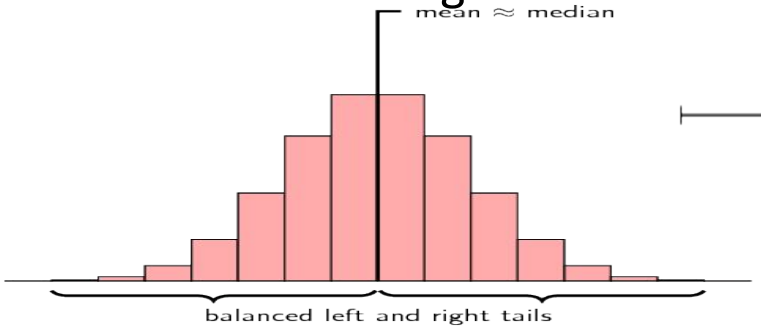
- Variance and standard deviation (*sample: s, population: σ)*
  - Variance: (algebraic, scalable computation)

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right] \qquad \sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 \qquad \frac{1}{N}\sum_{i=1}^{n} x_i^2 - \mu^2$$

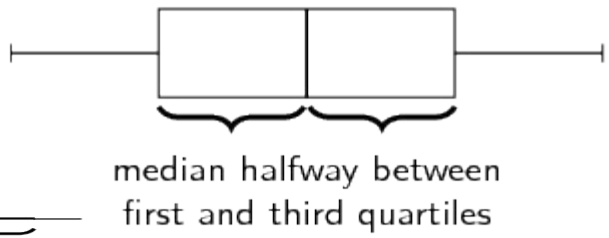
- Standard deviation s (or σ) is the square root of variance s2 (or σ2)
- Low standard deviation tells us that fewer numbers are far away from the mean.
- High standard deviation tells us that more numbers are far away from the mean.

Data Distribution representation using Histogram

- Normal Distribution
- Symmetric and skewed data (EMBKD)
- Distribution where the left and right hand sides of the distribution are roughly equally balanced around the mean
- The mean is approximately equal to the median
- Tails of the distribution are the parts to the left and to the right, away from the mean (counts in the histogram become smaller)
  - the left and right tails are equally balanced, meaning that they have about the same length



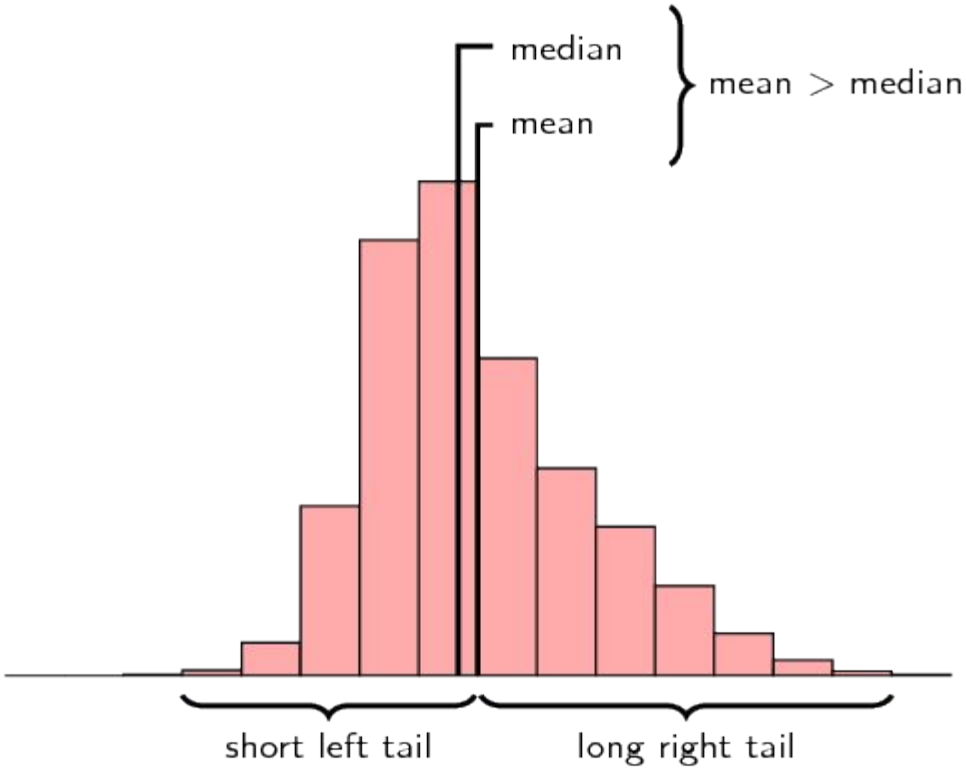
Histogram

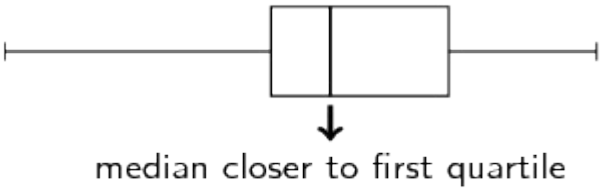


Box plot (Whisker diagram)

Skewed data distribution

- A distribution that is **skewed right** (also known as **positively skewed**) is shown below.

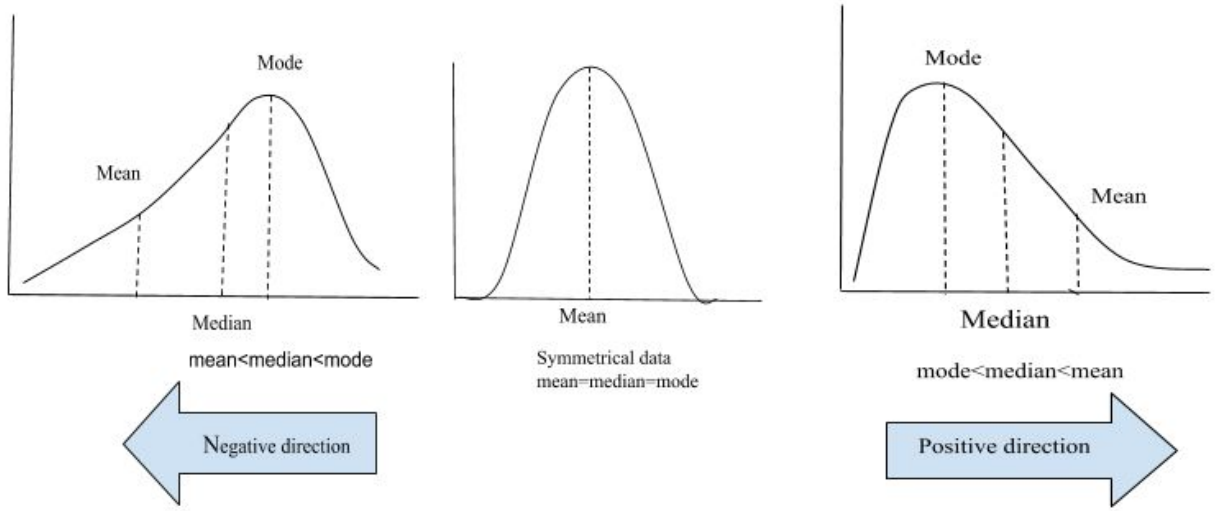




Skewed Data Distribution

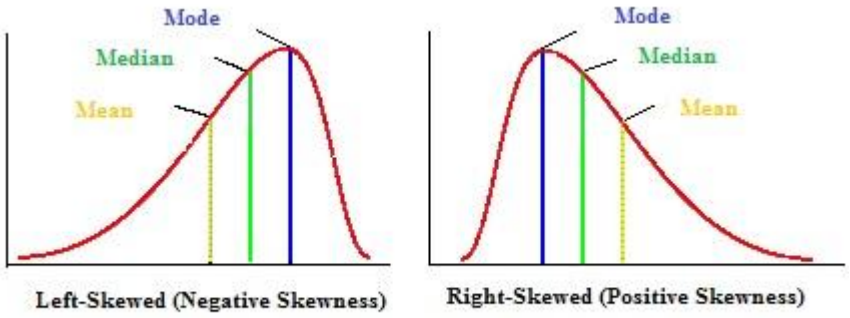
- Skewed Left Distribution :
  - has exactly the opposite characteristics of one that is skewed right:
  - the mean is typically less than the median;
  - the tail of the distribution is longer on the left hand side than on the right hand side; and
  - the median is closer to the third quartile than to the first quartile.

Skewed Data distribution



- The mean, median and mode are all Centrality measures (center of a set of data).
- The skewness of the data can be determined by how these quantities are related to one another
- By studying the shape of the data we can discover the relation between the mean, median and mode
- If the mean > median it indicates that the distribution is positively skewed. If the mean is < median it indicates that the distribution is negatively skewed

Skewness



$$S_K = \frac{mean - mode}{standard\ deviation}$$

Kal Pearson's first method uses mode and it's formula

If SK = 0 then we can say that the frequency distribution is normal and symmetrical.
If SK < 0 then we can say that the frequency distribution is negatively skewed.
If SK > 0 then we can say that the frequency distribution is positively skewed.

Measuring the Dispersion of Data (Detail with Box Plot theory)

- Quartiles, outliers and boxplots
  - Quartiles: Q_1 (25^{th} percentile), Q_3 (75^{th} percentile)
  - Inter-quartile range: IQR = $Q_3 - Q_1$
  - Five number summary: min, Q_1, M, Q_3, max
  - Boxplot: ends of the box are the quartiles, median is marked, whiskers, and plot outlier individually
  - Outlier: usually, a value higher/lower than 1.5 x IQR

Data Conversion

- Some tools can deal with nominal values but other need fields to be numeric
- Convert ordinal fields to numeric to be able to use ">" and "<" comparisons on such fields.
  - A 4.0
  - A- 3.7
  - B+ 3.3
  - B 3.0
- Multi-valued, unordered attributes with small no. of values
  - e.g. Color=Red, Orange, Yellow, …, Violet
  - for each value *v* create a binary "flag" variable *C_v* , which is 1 if Color=*v*, 0 otherwise

Conversion: Nominal, Many Values

- Examples:
  - US State Code (50 values)
  - Profession Code (7,000 values, but only few frequent)
- Ignore ID-like fields whose values are unique for each record
- For other fields, group values "naturally":
  - e.g. 50 US States 3 or 5 regions
  - Profession – select most frequent ones, group the rest
- Create binary flag-fields for selected values

Data Cleaning

- Data cleaning tasks:
  - Fill in missing values
  - Identify outliers and smooth out noisy data
  - Correct inconsistent data
  - Resolve redundancy caused by data integration

Missing Data

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not register history or changes of the data
- Missing data may need to be inferred.

How to Handle Missing Data?

- (1). Ignore the tuple (record) : usually done when class label is missing (assuming the tasks in classification)
- It is not effective when the percentage of missing values per attribute varies considerably.
- (2) Fill in the missing value manually: tedious + infeasible?

How to Handle Missing Data?

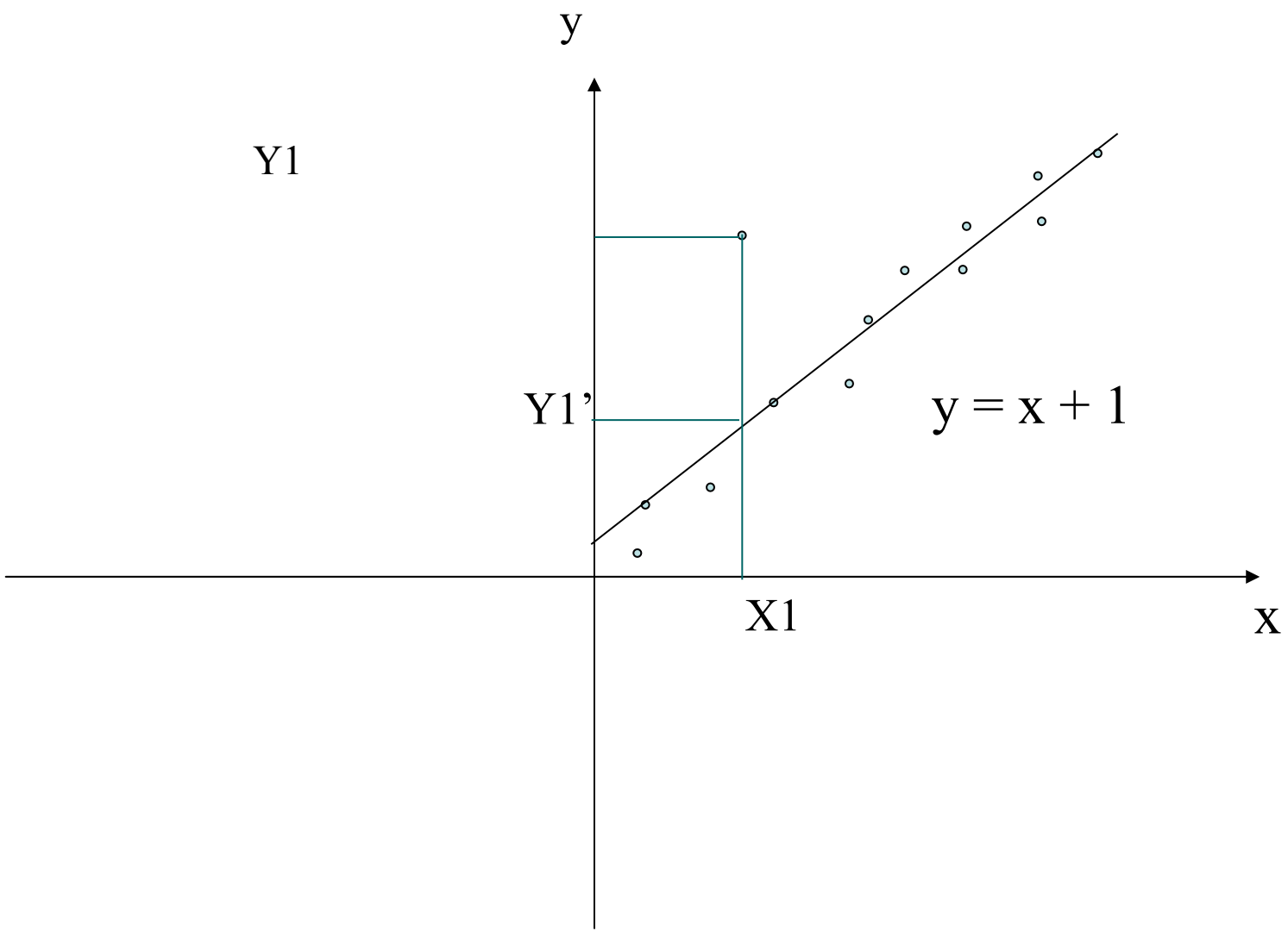
- (3) Use a global constant to fill in the missing value (introduces a new class)
- (4) Use the attribute values mean to fill in the missing value
- (5) Use the attribute values mean for all samples belonging to the same class to fill in the missing value: smarter then (4) in case of classification
- (6) Use the most probable value to fill in the missing value
- (7) Use regresion methods

Regression and Log-Linear Models

- Linear regression: Data are modeled to fit a straight line
  - Often uses the least-square method to fit the line
- Multiple regression: allows a response variable Y to be modeled as a linear function of multidimensional feature vector
- Log-linear model: approximates discrete multidimensional probability distributions

Linear Regression

Use regression analysis on values of an attributes to fill missing values.



Regression and Log-Linear Models

- Linear regression: $Y = \alpha + \beta X$
  - Two parameters , α and β specify the line and are to be estimated by using the data at hand.
  - using the least squares criterion to the known values of

$Y_1, Y_2, \ldots, X_1, X_2, \ldots$

- Multiple regression: $Y = b0 + b1\ X1 + b2\ X2.$
  - Many nonlinear functions can be transformed into the above.
- Log-linear models:
  - The multi-way table of joint probabilities is approximated by a product of lower-order tables.
  - Probability: $p(a, b, c, d) = \alpha_{ab}\ \beta_{ac} \chi_{ad}\ \delta_{bcd}$