Data Pre-Processing:

— UNIT-02 —

Jiawei Han, Micheline Kamber, and Jian Pei
University of Illinois at Urbana-Champaign &
Simon Fraser University

Why Data Mining?

- The Explosive Growth of Data
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: Remote sensing, bioinformatics, scientific simulation, …
    - Society and everyone: news, digital cameras, YouTube
- We are drowning in data, but starving for knowledge!
- "Necessity is the mother of invention"—Data mining—Automated analysis of massive data sets

What Is Data Mining?

- Data mining (knowledge discovery from data)
  - Extraction of interesting patterns or knowledge from huge amount of data
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

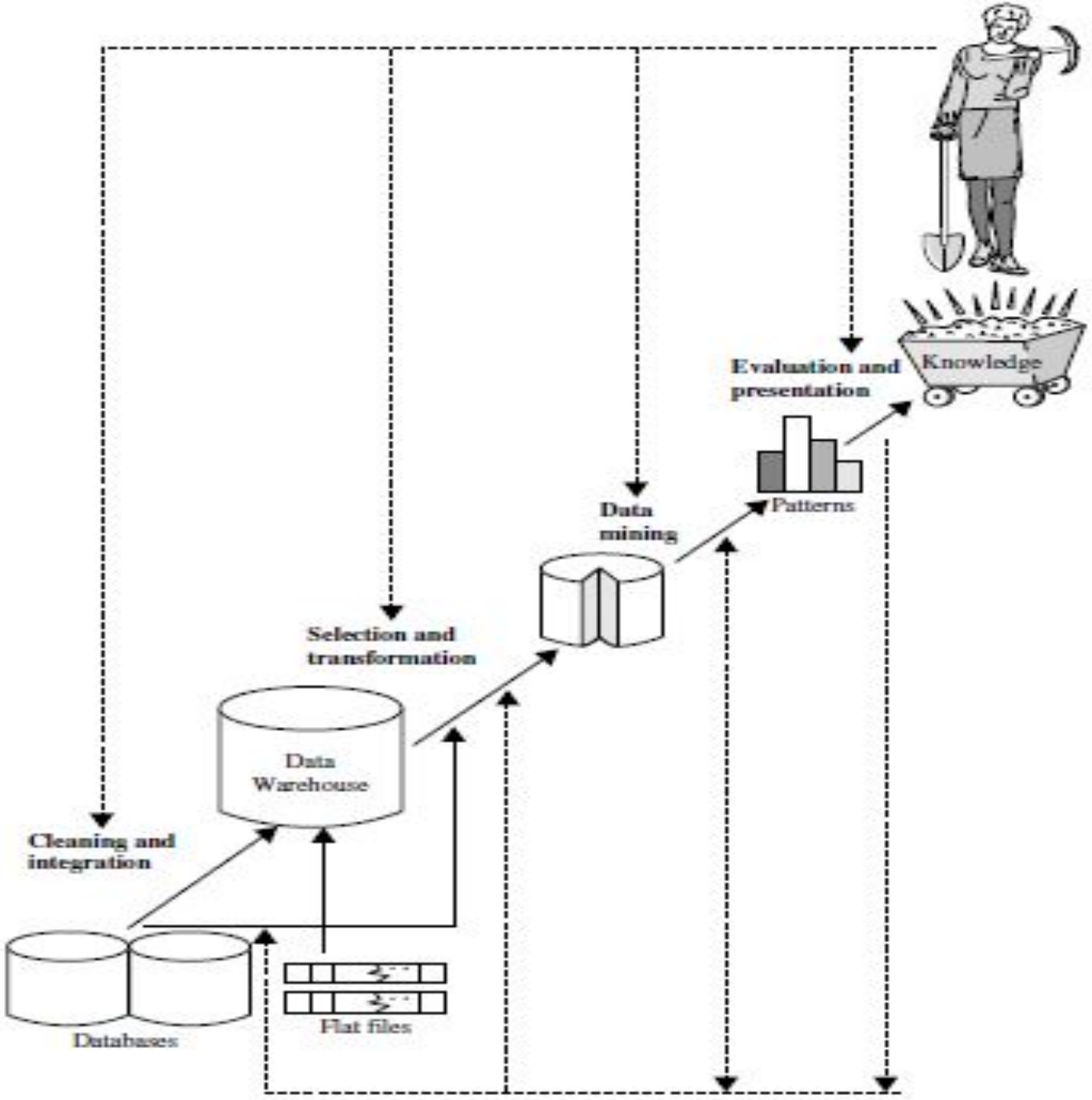


Figure 1.4 Data mining as a step in the process of knowledge discovery.

- **1. Data cleaning** (to remove noise and inconsistent data)
- **2. Data integration** (where multiple data sources may be combined)
- **3. Data selection** (where data relevant to the analysis task are retrieved from the database)
- **4. Data transformation** (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations)
- **5. Data mining** (an essential process where intelligent methods are applied to extract data patterns)
- **6. Pattern evaluation** (to identify the truly interesting patterns representing knowledge*)*
- **7. Knowledge presentation** (where visualization and knowledge representation techniques are used to present mined knowledge to users)

- Low-quality data will lead to low-quality mining results
- Data processing techniques, when applied before mining, can substantially improve the overall quality of the patterns mined

Data Quality: Why Preprocess the Data?

Measures for data quality: A multidimensional view

- Accuracy: correct or wrong, accurate or not
- Completeness: not recorded, unavailable, …
- Consistency: some modified but some not, dangling, …
- Timeliness: timely update?
- Believability: how trustable the data are correct?
- Interpretability: how easily the data can be understood?

Major Tasks in Data Preprocessing

- **Data cleaning**
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- **Data integration**
  - Integration of multiple databases, files etc
- **Data reduction**
  - Dimensionality reduction
  - Numerosity reduction
  - Data compression
- **Data transformation and data discretization**
  - Normalization
  - Concept hierarchy generation

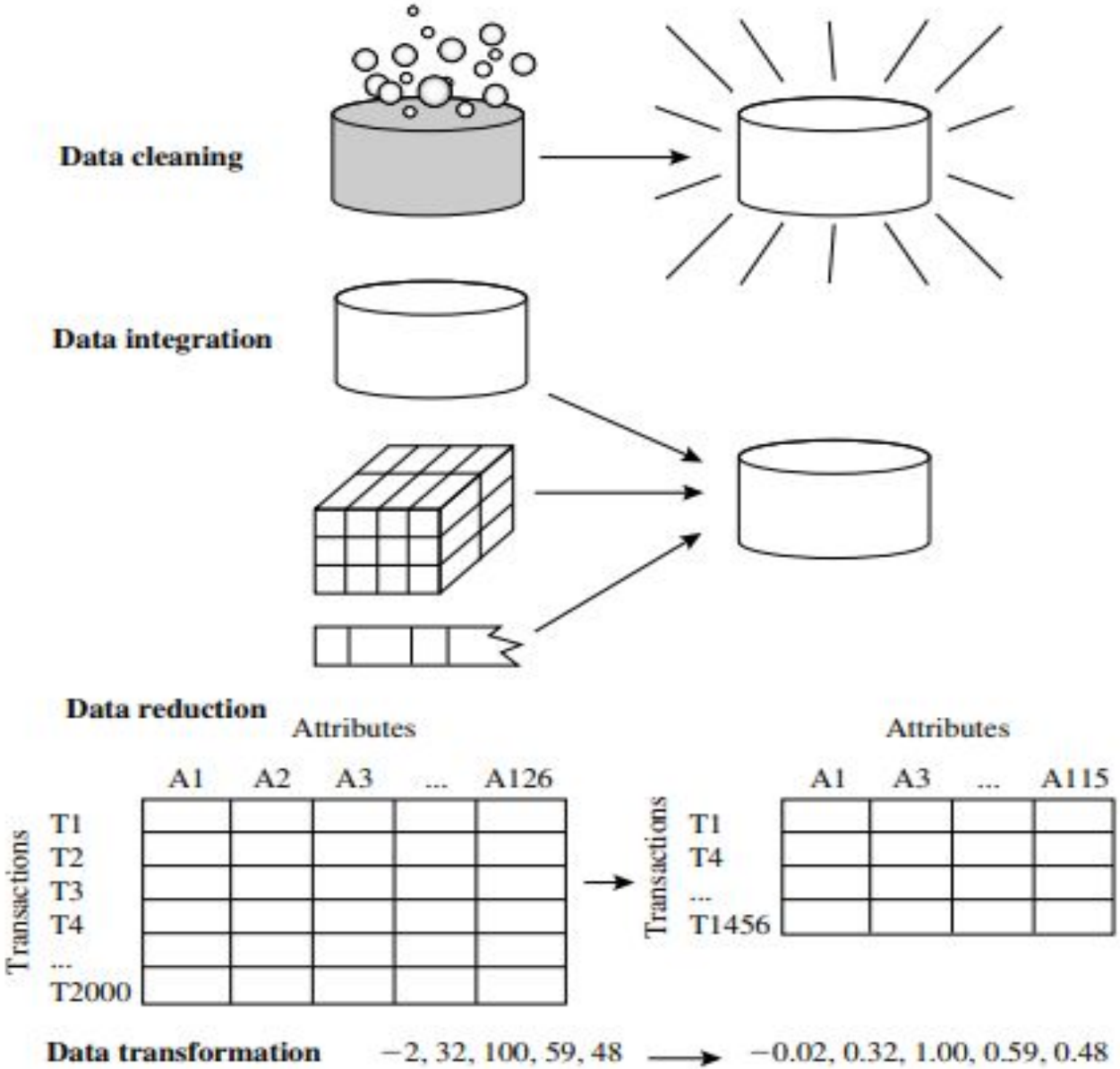


Figure 3.1 Forms of data preprocessing.

Data Cleaning

- Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, transmission error
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., *Occupation*=“ ” (missing data)
  - noisy: containing noise, errors, or outliers
    - e.g., *Salary*=“−10” (an error)
  - inconsistent: containing discrepancies in codes or names, e.g.,
    - *Age*=“42”, *Birthday*=“03/07/2010”
    - Was rating “1, 2, 3”, now rating “A, B, C”
    - discrepancy between duplicate records

Incomplete (Missing) Data

- Data is not always available
- 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?

- **Ignore the tuple**: usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably
- **Fill in the missing value manually:** tedious + infeasible?
- **Use a global constant to fill in the missing value:**
  - Replace all missing attribute values by the same constant such as a label like “Unknown” or −∞.
  - If missing values are replaced by, say, “Unknown,” then the mining program may mistakenly think that they form an interesting concept, since they all have a value in common—that of “Unknown.”
  - Hence, although this method is simple, it is not foolproof.

- **Use a measure of central tendency for the attribute to fill in the missing value:**
  - For example, suppose that the data distribution regarding the income of AllElectronics customers is symmetric and that the mean income is $56,000. Use this value to replace the missing value for income
- **Use the attribute mean or median for all samples belonging to the same class as the given tuple**
  - Replace the missing value with mean/median value of of that class

- **Use the most probable value to fill in the missing value:**
  - This may be determined with regression, inference-based tools using a Bayesian formalism, or decision tree

Noisy Data

- **Noise**: random error or variance in a measured variable
- outliers, which may represent noise
- **Data Smoothing Techniques**
- Binning methods smooth a sorted data value by consulting its "neighborhood," that is, the values around it.
- The sorted values are distributed into a number of "buckets," or bins.
- Because binning methods consult the neighborhood of values, they perform local smoothing

Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:

Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing by bin means:

Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

Smoothing by bin boundaries:

Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

Figure 3.2 Binning methods for data smoothing.

How to Handle Noisy Data?

- Regression
  - smooth by fitting the data into regression functions
- Clustering
  - detect and remove outliers
- Combined computer and human inspection
  - detect suspicious values and check by human (e.g., deal with possible outliers)

- Given is the frequency of stop words in documents (The values are given in increasing order) :
- 40, 45, 46, 13, 15, 16, 16, 19, 20, 20, 25, 25, 30, 33,33, 35, 35, 35, 35, 36, 52, 70, 21, 22, 22, 25, 25
- (i) Use smoothing by bin means with a depth of 3.

Data Integration

- **Data integration**:
  - Combines data from multiple sources into a coherent store
  - Careful integration can help reduce and avoid redundancies and inconsistencies in the resulting data set.
  - This can help improve the accuracy and speed of the subsequent data mining process.
- There are a number of **issues** to consider during data integration.
  - Schema integration
  - Object matching can be tricky.

- Schema integration: e.g., A.cust-id ≡ B.custumer_ID
  - Integrate metadata from different sources
- Entity identification problem:
  - Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton
- Detecting and resolving data value conflicts
  - For the same real world entity, attribute values from different sources are different
  - Possible reasons: different representations, different scales, e.g., Rs vs. Dollar

Handling Redundancy in Data Integration

- Redundant data occur often when integration of multiple databases
  - *Object identification*: The same attribute or object may have different names in different databases
  - *Derivable data:* One attribute may be a "derived" attribute in another table, e.g., annual revenue
- Redundant attributes may be able to be detected by *correlation analysis* and *covariance analysis*
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

How to Analyze Nominal Data?

Nominal data can be analyzed using the grouping method. The variables can be grouped together into categories, and for each category, the frequency or percentage can be calculated. The data can also be presented visually, such as by using a pie chart.

Although nominal data cannot be treated using mathematical operators, they still can be analyzed using advanced statistical methods. For example, one way to analyze the data is through hypothesis testing.

For nominal data, hypothesis testing can be carried out using nonparametric tests such as the chi-squared test. The chi-squared test aims to determine whether there is a significant difference between the expected frequency and the observed frequency of the given values.

χ 2(Chi-square) Correlation Test for Nominal Data

- For nominal data, a correlation relationship between two attributes, A and B, can be discovered by a χ 2 (chi-square) test.

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}, \tag{3.1}$$

where o_{ij} is the *observed frequency* (i.e., actual count) of the joint event (A_i, B_j) and e_{ij} is the *expected frequency* of (A_i, B_j), which can be computed as

$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{n}, \tag{3.2}$$

where n is the number of data tuples, $count(A = a_i)$ is the number of tuples having value a_i for A, and $count(B = b_j)$ is the number of tuples having value b_j for B. The sum in Eq. (3.1) is computed over all of the $r \times c$ cells. Note that the cells that contribute the most to the χ^2 value are those for which the actual count is very different from that expected.

The χ^2 statistic tests the hypothesis that A and B are *independent*, that is, there is no correlation between them. The test is based on a significance level, with $(r - 1) \times (c - 1)$ degrees of freedom. We illustrate the use of this statistic in Example 3.1. If the hypothesis can be rejected, then we say that A and B are statistically correlated.

Table 3.1 Example 2.1's 2 × 2 Contingency Table Data

	male	*female*	*Total*
fiction	250 (90)	200 (360)	450
non_fiction	50 (210)	1000 (840)	1050
Total	300	1200	1500

Note: Are *gender* and *preferred_reading* correlated?

Using Eq. (3.2), we can verify the expected frequencies for each cell. For example, the expected frequency for the cell (*male, fiction*) is

$$e_{11} = \frac{count(male) \times count(fiction)}{n} = \frac{300 \times 450}{1500} = 90,$$

and so on. Notice that in any row, the sum of the expected frequencies must equal the total observed frequency for that row, and the sum of the expected frequencies in any column must also equal the total observed frequency for that column.

Using Eq. (3.1) for χ^2 computation, we get

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840}$$

$$= 284.44 + 121.90 + 71.11 + 30.48 = 507.93.$$

For this 2×2 table, the degrees of freedom are $(2-1)(2-1) = 1$. For 1 degree of freedom, the χ^2 value needed to reject the hypothesis at the 0.001 significance level is 10.828 (taken from the table of upper percentage points of the χ^2 distribution, typically available from any textbook on statistics). Since our computed value is above this, we can reject the hypothesis that *gender* and *preferred_reading* are independent and conclude that the two attributes are (strongly) correlated for the given group of people. ■

Chi-square (χ^2)

Table of Observed Values

Qualification / Marital Status	Middle School	High School	Bachelor's	Master's	Ph.D	Total
Never married	18	36	21	9	6	90
Married	12	36	45	36	21	150
Divorced	6	9	9	3	3	30
Widowed	3	9	9	6	3	30
Total	39	90	84	54	33	300

Critical values of the Chi-square distribution with *d* degrees of freedom

	Probability of exceeding the critical value						
d	0.05	0.01	0.001	*d*	0.05	0.01	0.001
1	3.841	6.635	10.828	11	19.675	24.725	31.264
2	5.991	9.210	13.816	12	21.026	26.217	32.910
3	7.815	11.345	16.266	13	22.362	27.688	34.528
4	9.488	13.277	18.467	14	23.685	29.141	36.123
5	11.070	15.086	20.515	15	24.996	30.578	37.697
6	12.592	16.812	22.458	16	26.296	32.000	39.252
7	14.067	18.475	24.322	17	27.587	33.409	40.790
8	15.507	20.090	26.125	18	28.869	34.805	42.312
9	16.919	21.666	27.877	19	30.144	36.191	43.820
10	18.307	23.209	29.588	20	31.410	37.566	45.315

Covariance of numeric data

- In probability theory and statistics, correlation and covariance are two similar measures for assessing how much two attributes change together
- Consider two numeric attributes A and B, and a set of n observations {(a1,b1),...,(an,bn)}.
- The mean values of A and B, respectively, are also known as the **expected values** on A and B, that is

$$E(A) = \bar{A} = \frac{\sum_{i=1}^{n} a_i}{n}$$

and

$$E(B) = \bar{B} = \frac{\sum_{i=1}^{n} b_i}{n}.$$

Covariance (Numeric Data)

- Covariance is similar to correlation

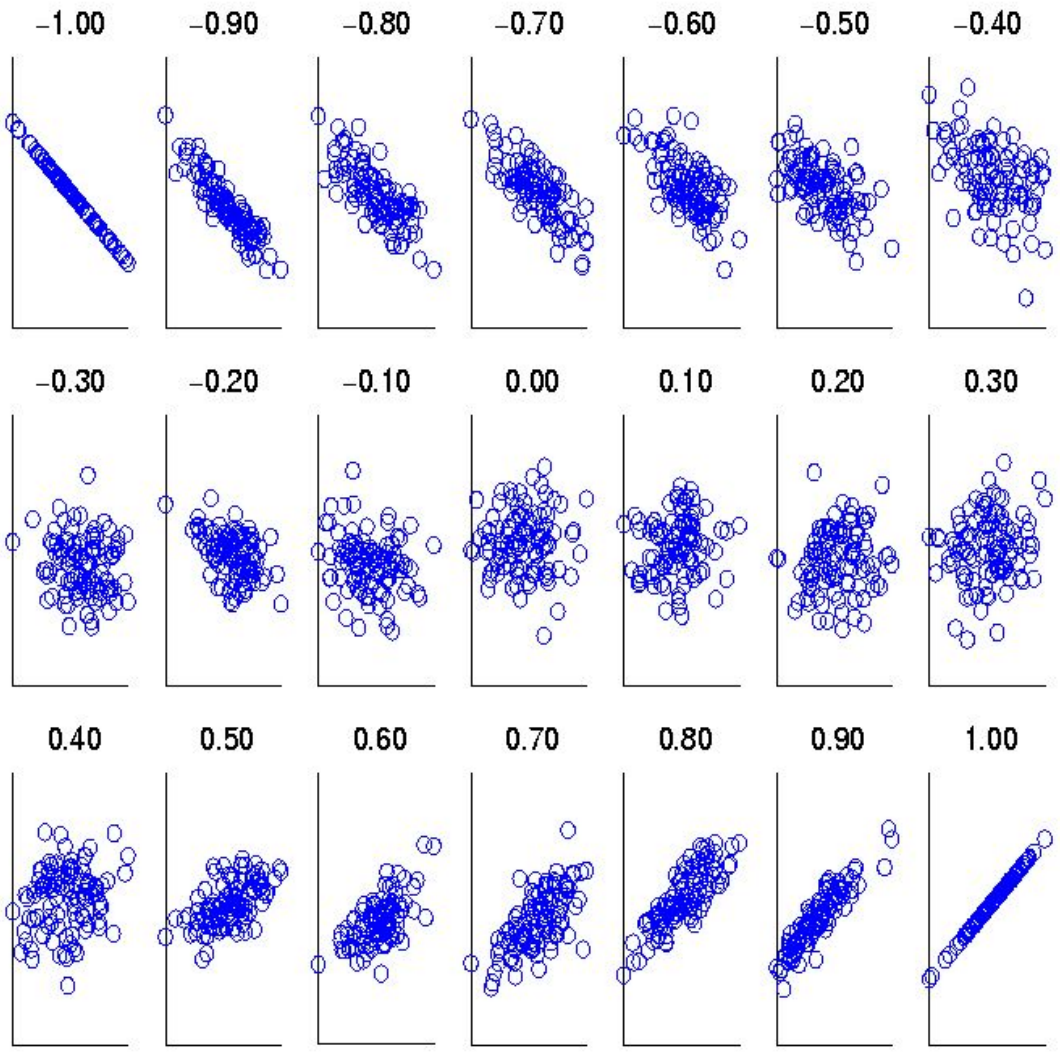
$$Cov(A,B) = E((A-\bar{A})(B-\bar{B})) = \frac{\sum_{i=1}^{n}(a_i-\bar{A})(b_i-\bar{B})}{n}$$

Correlation coefficient: $r_{A,B} = \frac{Cov(A,B)}{\sigma_A \sigma_B}$

where n is the number of tuples, $\bar{A}$ and $\bar{B}$ are the respective mean or **expected values** of A and B, σ_A and σ_B are the respective standard deviation of A and B.

- **Positive covariance**: If $Cov_{A,B} > 0$, then A and B both tend to be larger than their expected values.
- **Negative covariance**: If $Cov_{A,B} < 0$ then if A is larger than its expected value, B is likely to be smaller than its expected value.
- **Independence**: $Cov_{A,B} = 0$ but the converse is not true:

Visually Evaluating Correlation



Scatter plots showing the similarity from –1 to 1.

Co-Variance: An Example

$$Cov(A,B) = E((A-\bar{A})(B-\bar{B})) = \frac{\sum_{i=1}^{n}(a_i-\bar{A})(b_i-\bar{B})}{n}$$

- It can be simplified in computation as

$$Cov(A,B) = E(A \cdot B) - \bar{A}\bar{B}$$

- Suppose two stocks A and B have the following values in one week: (2, 5), (3, 8), (5, 10), (4, 11), (6, 14).
- Question: If the stocks are affected by the same industry trends, will their prices rise or fall together?
  - E(A) = (2 + 3 + 5 + 4 + 6)/ 5 = 20/5 = 4
  - E(B) = (5 + 8 + 10 + 11 + 14) /5 = 48/5 = 9.6
  - Cov(A,B) = (2×5+3×8+5×10+4×11+6×14)/5 – 4 × 9.6 = 4
- Thus, A and B rise together since Cov(A, B) > 0.

Table 3.2 Stock Prices for *AllElectronics* and *HighTech*

Time point	*AllElectronics*	*HighTech*
t1	6	20
t2	5	10
t3	4	14
t4	3	5
t5	2	5

(e.g., the data follow multivariate normal distributions) does a covariance of 0 imply independence.

Example 3.2 **Covariance analysis of numeric attributes.** Consider Table 3.2, which presents a simplified example of stock prices observed at five time points for *AllElectronics* and *HighTech*, a high-tech company. If the stocks are affected by the same industry trends, will their prices rise or fall together?

$$E(AllElectronics) = \frac{6+5+4+3+2}{5} = \frac{20}{5} = \$4$$

and

$$E(HighTech) = \frac{20+10+14+5+5}{5} = \frac{54}{5} = \$10.80.$$

Thus, using Eq. (3.4), we compute

$$Cov(AllElectroncis, HighTech) = \frac{6 \times 20 + 5 \times 10 + 4 \times 14 + 3 \times 5 + 2 \times 5}{5} - 4 \times 10.80$$

$$= 50.2 - 43.2 = 7.$$

Therefore, given the positive covariance we can say that stock prices for both companies rise together. ■

Variance is a special case of covariance, where the two attributes are identical (i.e., the covariance of an attribute with itself). Variance was discussed in Chapter 2.

Temperature(X)	No. Of Customers(Y)
97	14
86	11
89	9
84	9
94	15
74	7

Example to understand correlation and covariance

Data Reduction Strategies

- **Data reduction**: Obtain a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results
- Why data reduction? — A database/data warehouse may store terabytes of data. Complex data analysis may take a very long time to run on the complete data set.
- Data reduction strategies
  - Dimensionality reduction, e.g., remove unimportant attributes
    - Attribute subset selection
  - Numerosity reduction (some simply call it: Data Reduction)
    - Histograms,
    - Clustering,
    - Sampling
    - Data cube aggregation
  - Data compression

Attribute Subset Selection

- Redundant attributes
  - Duplicate much or all of the information contained in one or more other attributes
  - E.g., purchase price of a product and the amount of sales tax paid
- Irrelevant attributes
  - Contain no information that is useful for the data mining task at hand
  - E.g., students' ID is often irrelevant to the task of predicting students' GPA

Attribute Subset Selection

- The “best” (and “worst”) attributes are typically determined using tests of statistical significance, which assume that the attributes are independent of one another.
- Many other attribute evaluation measures can be used such as the *information gain measure*

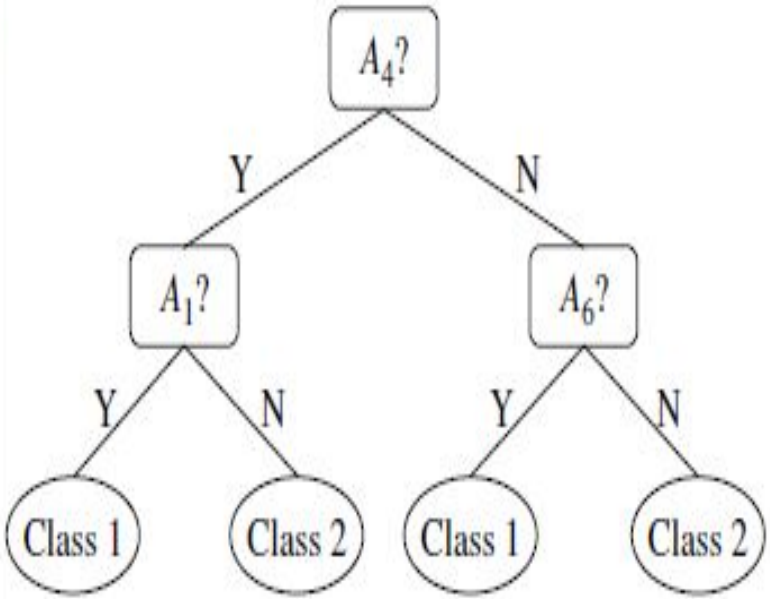
Forward selection	Backward elimination	Decision tree induction
Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ Initial reduced set: $\{\}$ $\Rightarrow \{A_1\}$ $\Rightarrow \{A_1, A_4\}$ $\Rightarrow$ Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ $\Rightarrow \{A_1, A_3, A_4, A_5, A_6\}$ $\Rightarrow \{A_1, A_4, A_5, A_6\}$ $\Rightarrow$ Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$  $\Rightarrow$ Reduced attribute set: $\{A_1, A_4, A_6\}$

Figure 3.6 Greedy (heuristic) methods for attribute subset selection.

Data Reduction 2: Numerosity Reduction

- Reduce data volume by choosing alternative, *smaller forms* of data representation
- **Histograms:**
- **It** use binning to approximate data distributions and are a popular form of data reduction.
- A **histogram for an** attribute, *A, partitions the data distribution of A into disjoint subsets, referred to as buckets or bins.*
- *If each bucket represents only a single attribute–value/frequency pair, the* buckets are called *singleton buckets.*
- *Often, buckets instead represent continuous ranges* for the given attribute.

Histograms. The following data are a list of A*llElectronics prices for commonly sold* items (rounded to the nearest dollar). The numbers have been sorted:

1, 1, 5, 5, 5, 5, 5, 8, 8, 10, 10, 10, 10, 12, 14, 14, 14, 15, 15, 15, 15, 15, 15, 18, 18, 18, 18, 18, 18, 18, 18, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30, 30, 30.

- Figure shows a histogram for the data using singleton buckets.
- To further reduce the data, it is common to have each bucket denote a continuous value range for the given attribute.
- In Figure ,each bucket represents a different $10 range for *price.*

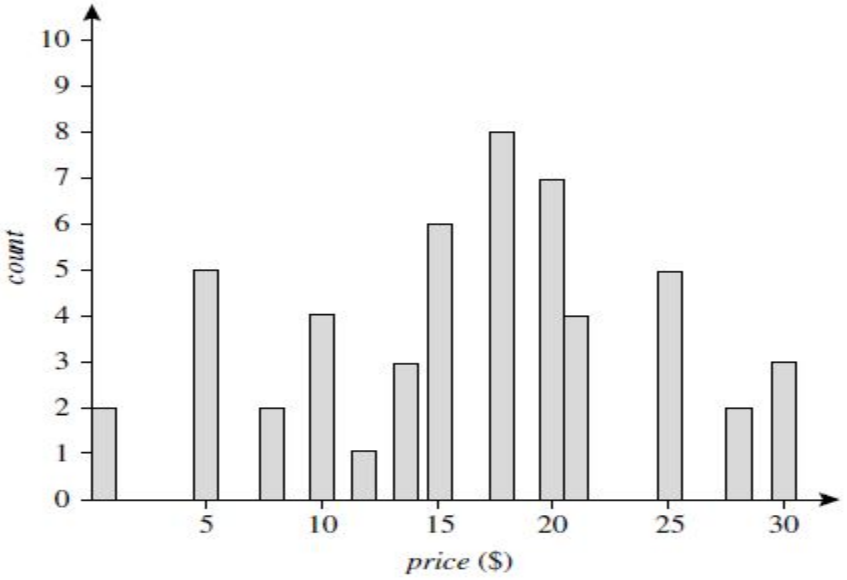


Figure 3.7 A histogram for *price* using singleton buckets—each bucket represents one price–value/

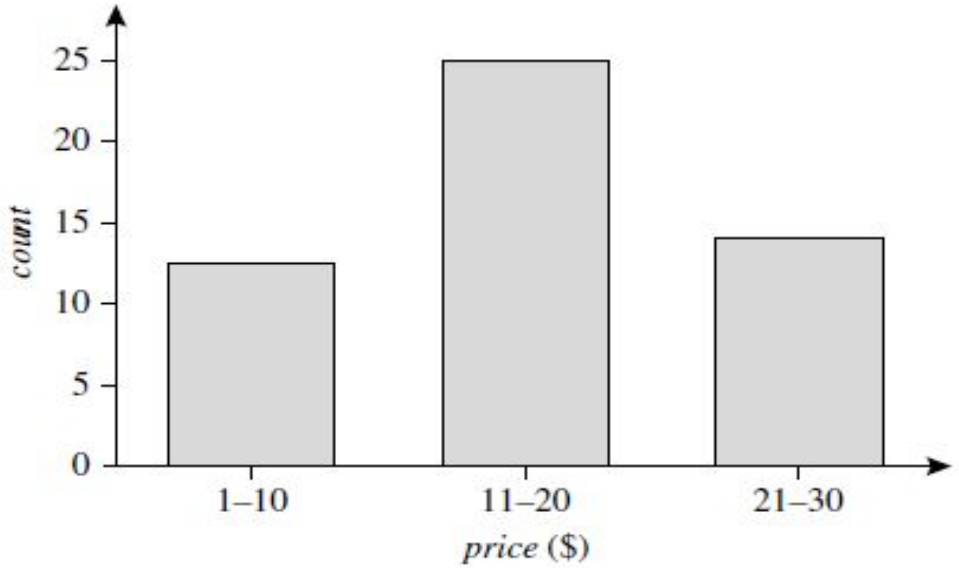


Figure 3.8 An equal-width histogram for *price*, where values are aggregated so that each bucket has a uniform width of $10.

- There are several partitioning rules,
- **Equal-width**: In an equal-width histogram, the width of each bucket range is uniform
- **Equal-frequency (or equal-depth):** In an equal-frequency histogram, the buckets are created so that, roughly, the frequency of each bucket is constant (i.e., each bucket

contains roughly the same number of contiguous data samples).

- Suppose a group of 12 *sales price* records has been sorted as follows:
- 5*;* 10*;* 11*;* 13*;* 15*;* 35*;* 50*;* 55*;* 72*;* 92*;* 204*;* 215*:*
- Partition them into three bins by each of the following methods.
- (a) equal-frequency partitioning
- (b) equal-width partitioning

- Answer:

(a) equal-frequency partitioning

- bin 1 5,10,11,13
- bin 2 15,35,50,55
- bin 3 72,92,204,215

(**b) equal-width partitioning**

- The width of each interval is (215 - 5)/=3 = 70.
- bin 1 5,10,11,13,15,35,50,55,72
- bin 2 92
- bin 3 204,215

Clustering

- Clustering techniques consider data tuples as objects.

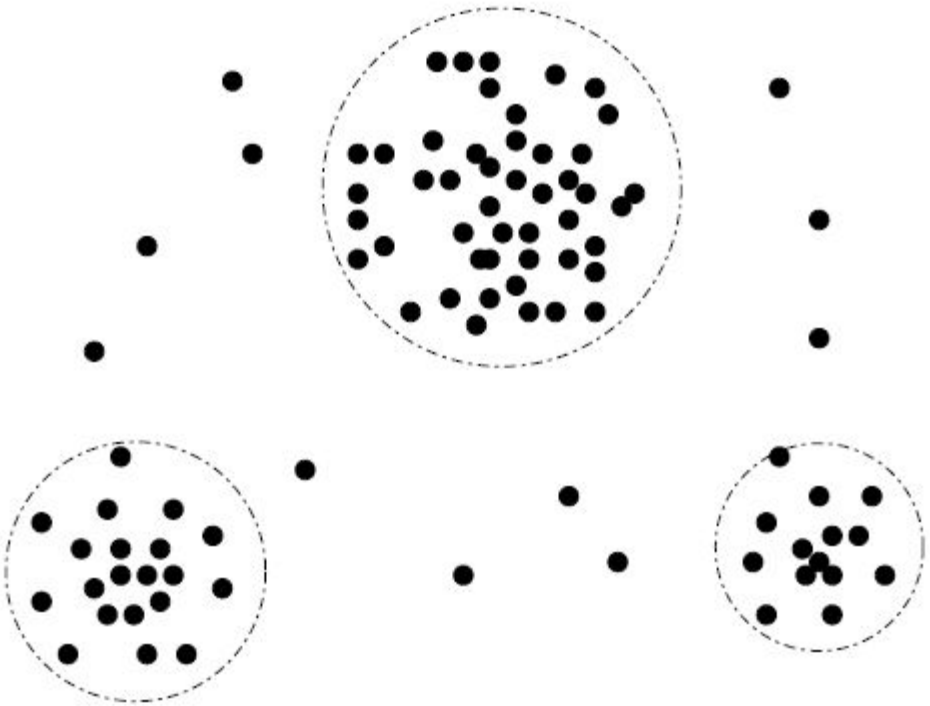
- Similarity is commonly defined in terms of how “close” the objects are in space, based on a distance function

- The “quality” of a cluster may be represented by **its diameter**, the maximum distance between any two objects in the cluster

- **Centroid distance** is an alternative measure of cluster quality and is defined as the average distance of each cluster object from the cluster centroid

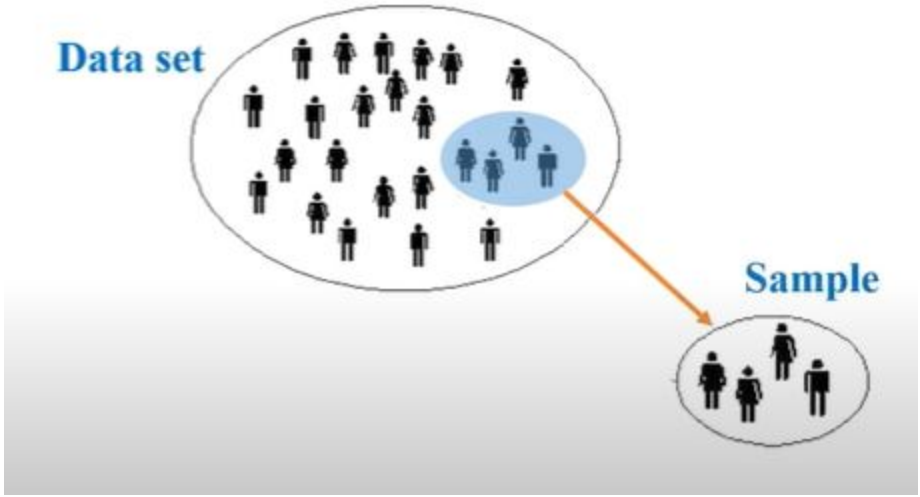
Clustering

- In data reduction, the cluster representations of the data are used to replace the actual data.
- The effectiveness of this technique depends on the data's nature.



Sampling

- Approach for selecting subset of data set for analysis
- Subset here is sample
- It is statistical Approach
- Sample should be representative



Consider that the property of interest is Mean of given data set.

{2, 7, 11, 12, 13, 15, 16}

Mean = 10.86

Let us find some sample out of the data given

Sample 1: (7, 12, 13) **Mean: 10.67**

Sample 2: (2, 15, 16) **Mean: 11**

Representative sample

Sample 3: (11, 13, 15) **Mean: 13**

Different Sampling Techniques include:

- Simple Random Sampling
- Stratified Sampling
- Progressive sampling- To identify sample size

Simple Random Sampling

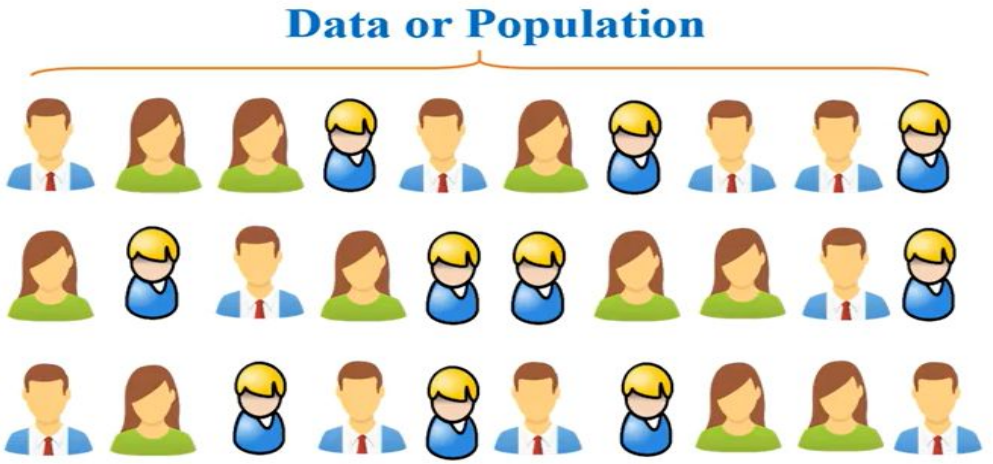
- Simplest of all sampling Techniques
- Equal opportunity of selection for all item in dataset

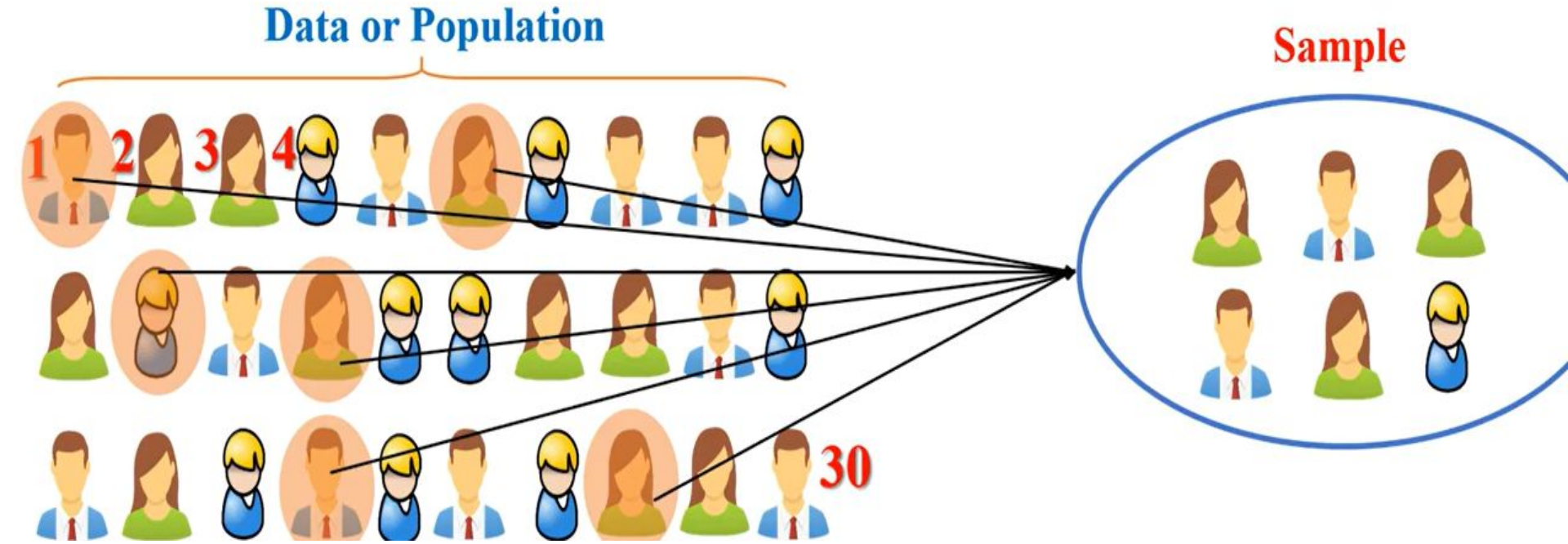
Two variation

- Sampling without Replacement
- Sampling with Replacement

Sampling without Replacement:-

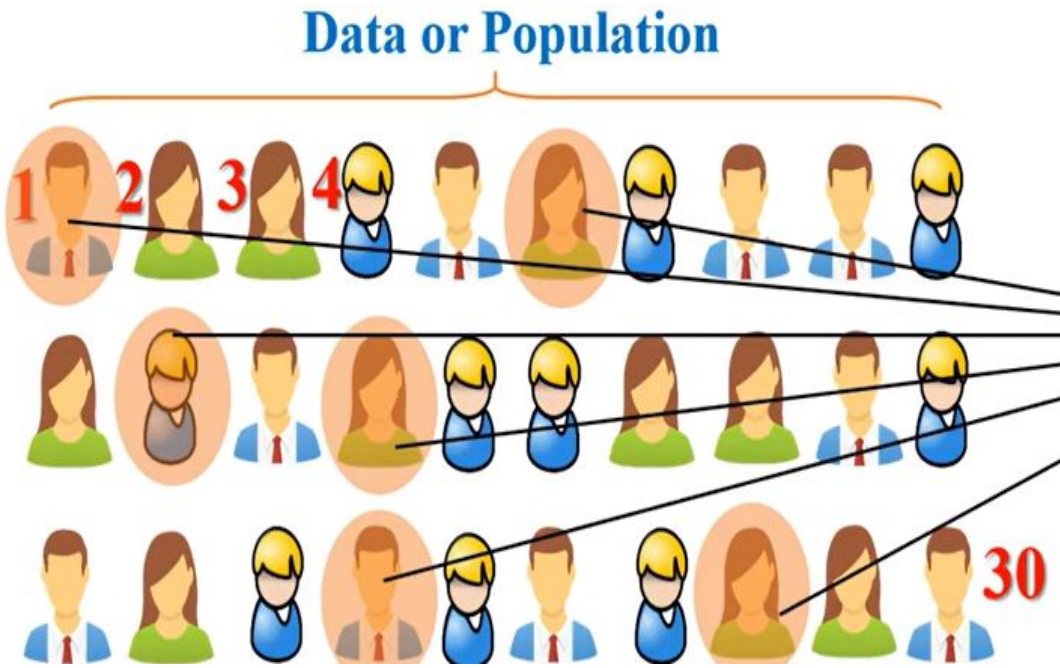
- Once item consider then it can not be considered as sample

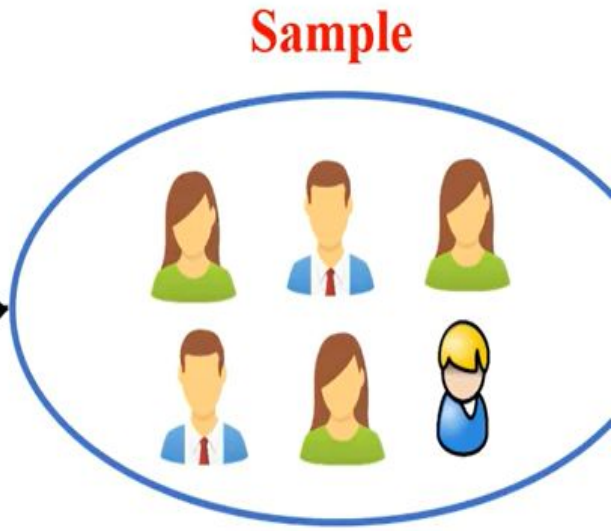


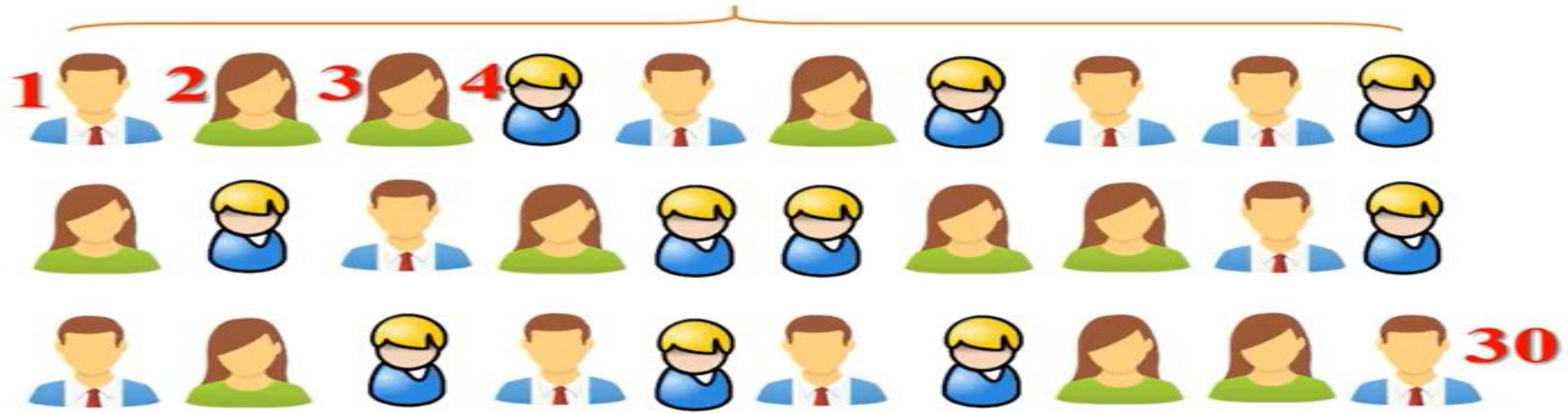

Data or Population
Sample
1
2
3
4
30


Data or Population after finding first sample
1
2
3
24

Sampling with Replacement:-

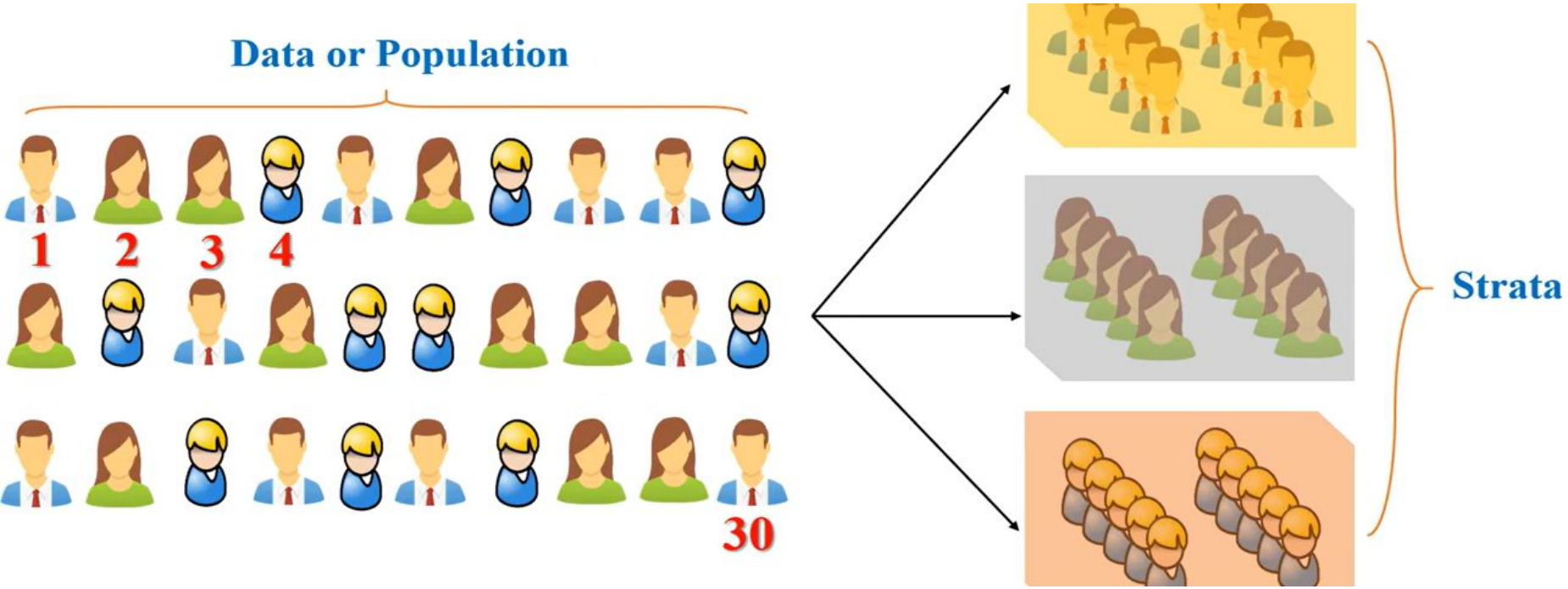




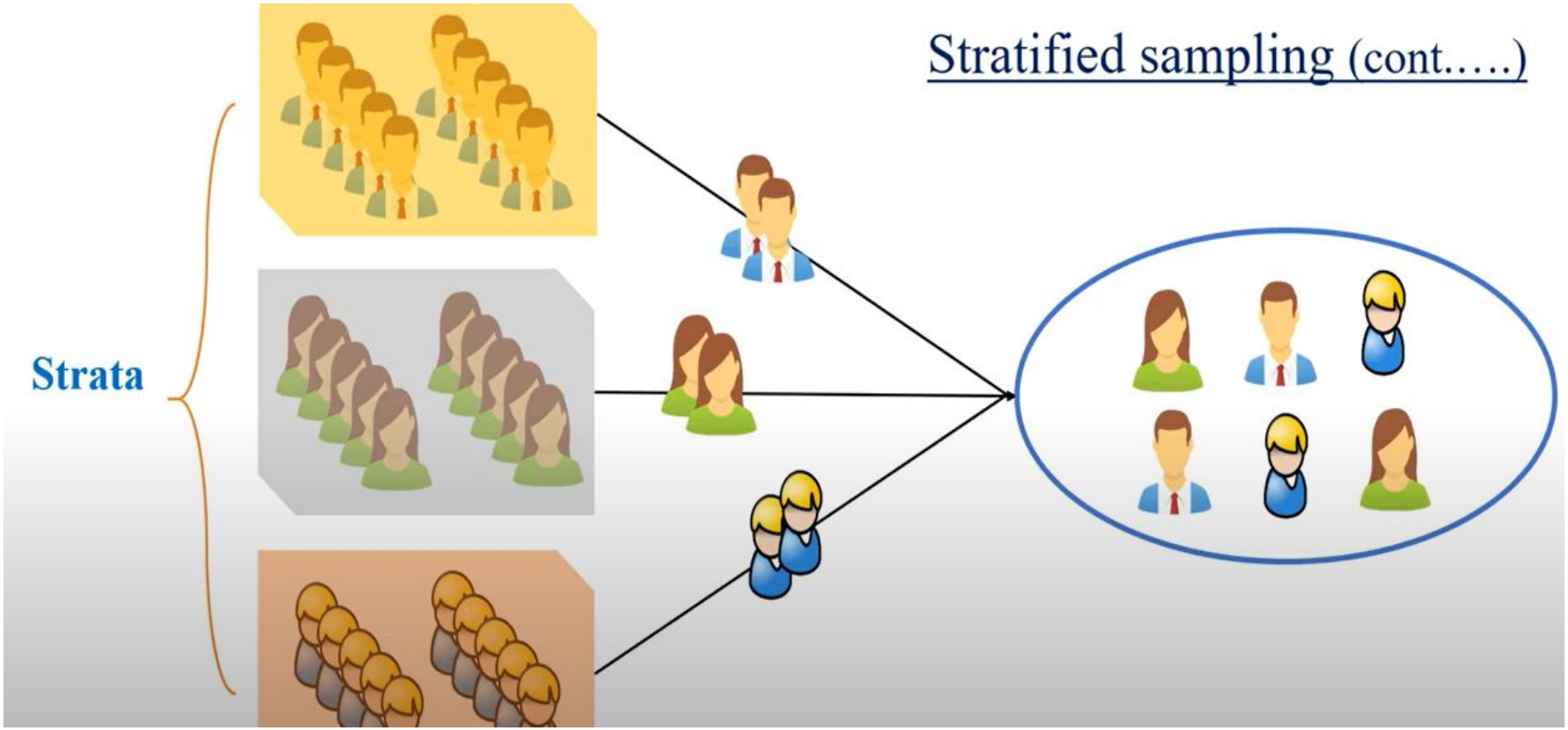


Stratified Sampling

- Start with pre-specified group of object
- Each group is called strata
- Group size may vary but sample size is fixed by data miner
- Selection of object for sample will be preferably same



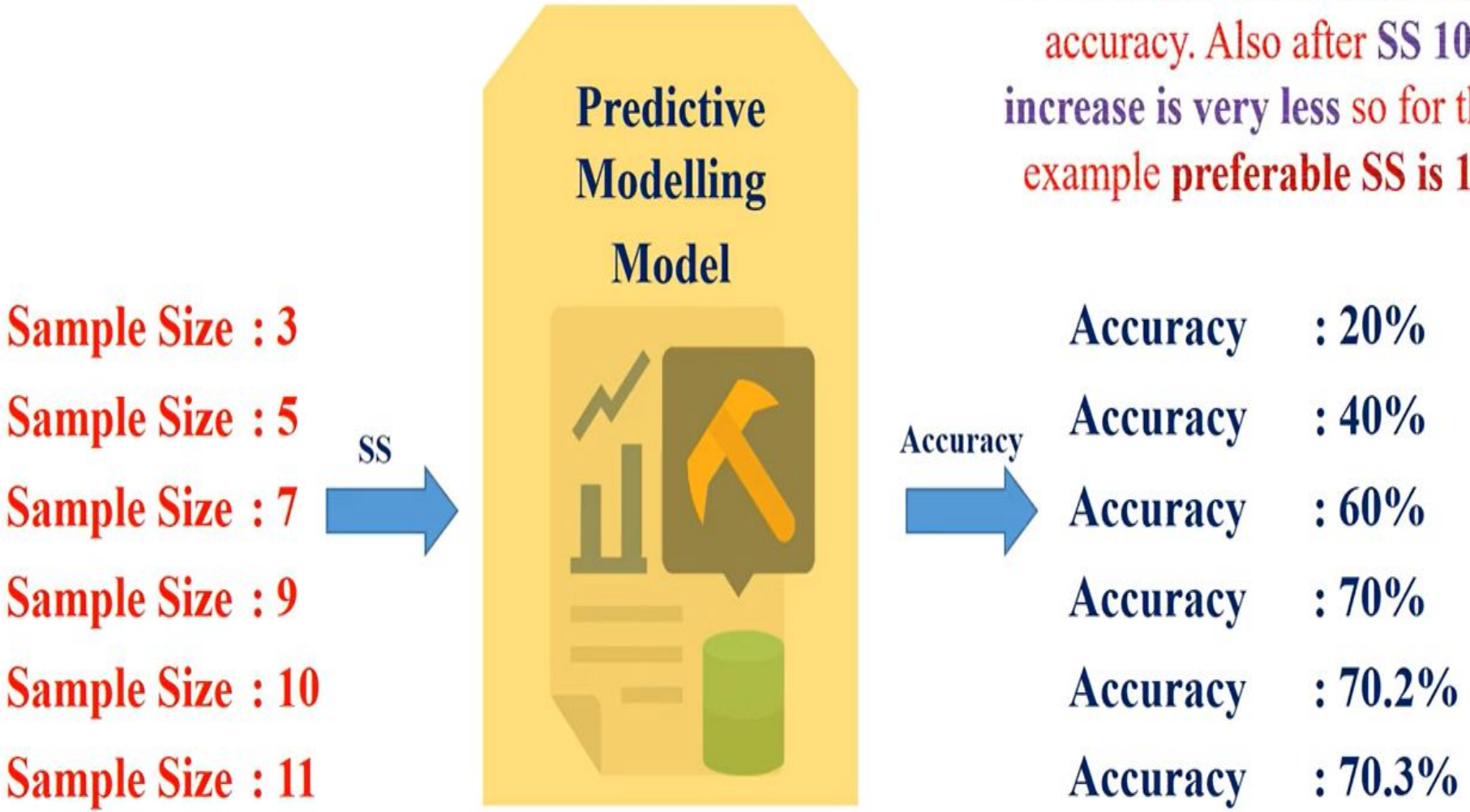
Stratified sampling (cont.….)



Progressive Sampling

Progressive sampling (cont.....)

Here, we can observe that after SS 9 there is small increase in accuracy. Also after SS 10 increase is very less so for this example preferable SS is 10



Types of Sampling

- **Simple random sampling**
  - There is an equal probability of selecting any particular item
  - **Sampling without replacement**
  - Once an object is selected, it is removed from the population
  - **Sampling with replacement**
  - A selected object is not removed from the population
- **Stratified sampling:**
  - Partition the data set, and draw samples from each partition (proportionally, i.e., approximately the same percentage of the data)

Data Cube Aggregation

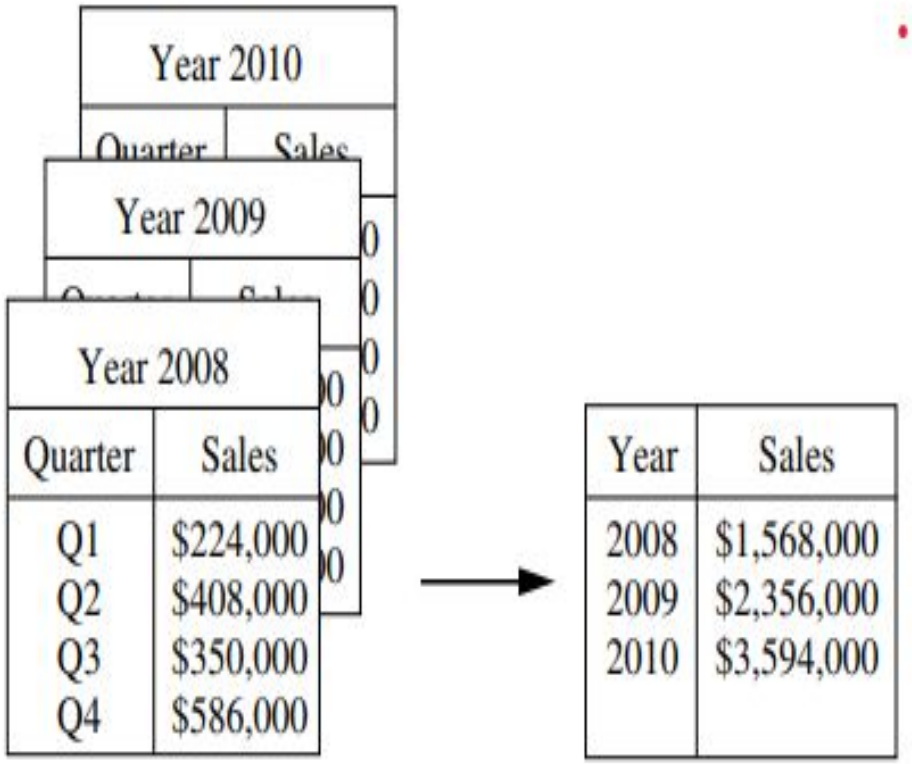
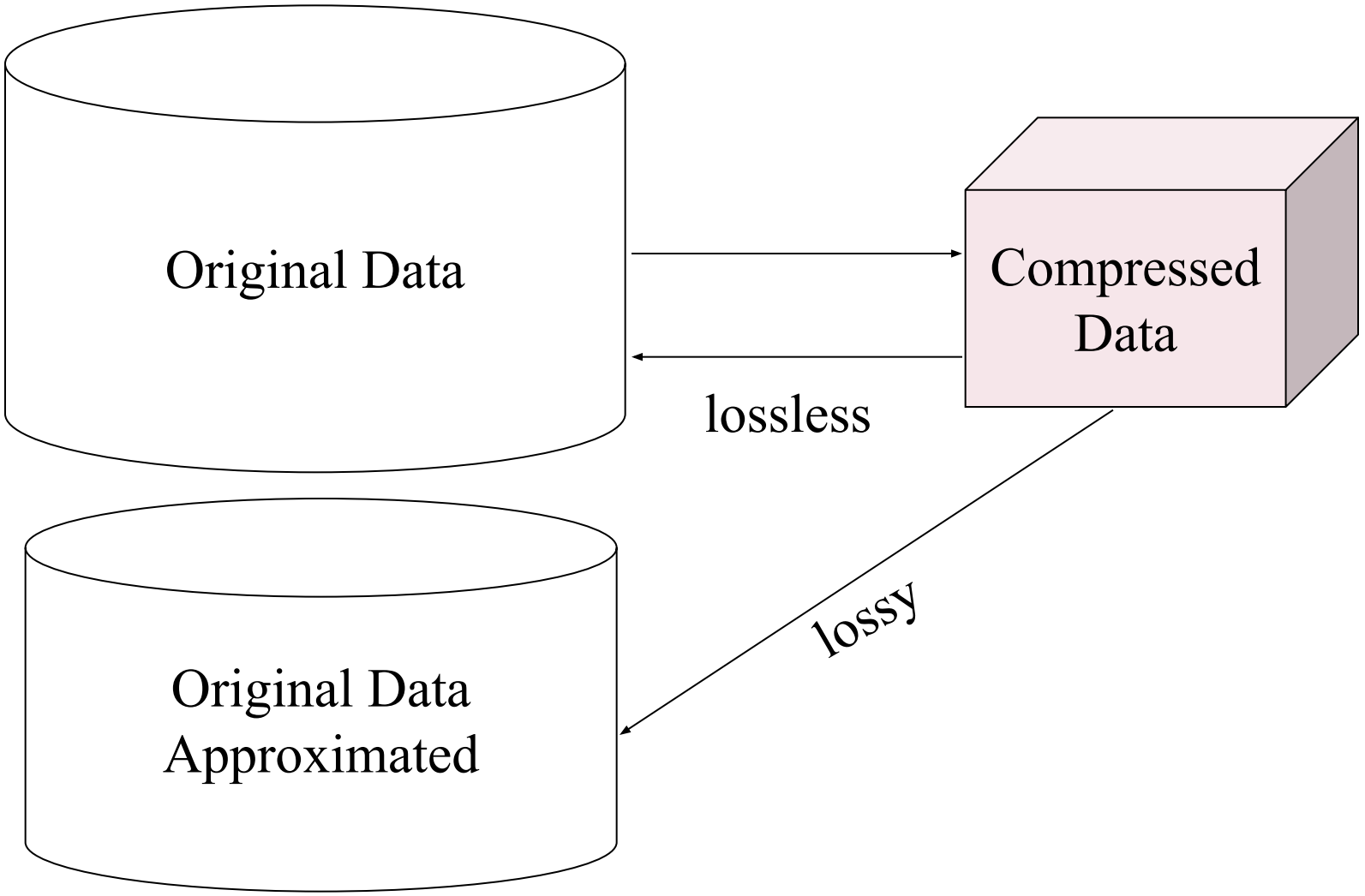


Figure 3.10 Sales data for a given branch of *AllElectronics* for the years 2008 through 2010. On the *left*, the sales are shown per quarter. On the *right*, the data are aggregated to provide the annual sales.

Data Compression



Data Transformation

- The data are transformed or consolidated so that the resulting mining process may be more efficient
- The patterns found after transformation may be easier to understand
- Methods
  - **Smoothing:** Remove noise from data
  - **Attribute/feature construction**
    - New attributes constructed from the given ones
  - **Aggregation:** Summarization, data cube construction

Normalization: Scaled to fall within a smaller, specified range

- min-max normalization
- z-score normalization
- normalization by decimal scaling

Discretization: Here the raw values of a numeric attribute are replaced by interval labels

Concept hierarchy generation

- The measurement unit used can affect the data analysis
- To help avoid dependence on the choice of measurement units, the data should be normalized or standardized.
- Normalizing the data attempts to give all attributes an equal weight
- **Min-max normalization**
- Performs a linear transformation on the original data.
- Suppose that minA and maxA are the minimum and maximum values of an attribute,
- A. Min-max normalization maps a value, vi , of A to vi’ in the range [new minA,new maxA] by computing vi’

Normalization

- **Min-max normalization**: to [new_min$_A$, new_max$_A$]

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,600 is mapped to

$$\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$$

- **Z-score normalization** :-
- In z-score normalization (or zero-mean normalization), the values for an attribute, A, are normalized based on the mean (i.e., average) and standard deviation of A.

$$v' = \frac{v - \mu_A}{\sigma_A}$$

$$\textit{Standard Deviation} = \sqrt{\frac{\sum(x_i - \mu)^2}{n-1}}$$

  - Ex. Let μ = 54,000, σ = 16,000 for income. Then z-score for 73600

$$\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$$

- **Normalization by decimal scaling**
- Normalizes by moving the decimal point of values of attribute A.
- The number of decimal points moved depends on the maximum absolute value of A.
- A value, vi , of A is normalized to vi' by computing

$$v_i' = \frac{v_i}{10^j},$$

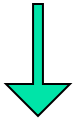
where j is the smallest integer such that $max(|v_i'|) < 1$.

- Suppose that the recorded values of A range from −986 to 917.
- The maximum absolute value of A is 986.
- To normalize by decimal scaling, we therefore divide each value by 1000 (i.e., j = 3)
- so that −986 normalizes to −0.986 and 917 normalizes to 0.917.

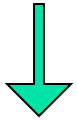
- Given is the frequency of stop words in documents (The values are given in increasing order) : 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33,33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70. Apply the following methods and show the results :—
- (i) Use smoothing by bin means with a depth of 3.
- (ii) Use Min – Max normalization to transform the value 30 into the range 0.0 to 1.0.
- (iii) Use z – score normalization to transform the value 30 where the standard deviation of the above frequency is 12.94.
- (iv) Use normalization by decimal scaling to transform the value 30.
- (v) Plot an equi – width histogram of width 10 on graph paper.

Discretization

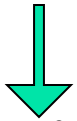
- Process of minimizing



- No of values



- A continuous attribute



- By diving it into interval

Discretization

- **Data Discretization**
  - **With Class**
    - Supervised
    - Unsupervised
  - **With Direction**
    - Top down
    - Bottom up

Data Discretization Methods with Numeric data

- Typical methods: All the methods can be applied recursively
  - Binning
  - Histogram analysis
  - Clustering analysis
  - Decision-tree analysis

Simple Discretization: Binning

- Equal-width (distance) partitioning
  - Divides the range into *N* intervals of equal size: uniform grid
  - if *A* and *B* are the lowest and highest values of the attribute, the width of intervals will be: $W = (B - A)/N$.
- Equal-depth (frequency) partitioning
  - Divides the range into *N* intervals, each containing approximately same number of samples

Binning Methods for Data Smoothing

- Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34, with 3 bin

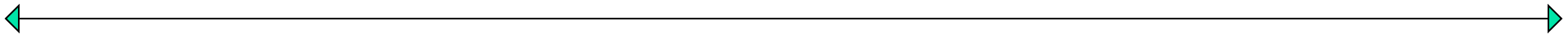
* Partition into equal-frequency (**equi-depth**) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
* Smoothing by **bin means**:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
* Smoothing by **bin boundaries**:
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

3. Exercise 2.2 gave the following data (in increasing order) for the attribute *age*: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

   (a) Use *smoothing by bin means* to smooth the above data, using a bin depth of 3. Illustrate your steps. Comment on the effect of this technique for the given data.

   (b) How might you determine *outliers* in the data?

   (c) What other methods are there for *data smoothing*?



- **Step 1:** Sort the data. (This step is not required here as the data are already sorted.)
- **Step 2:** Partition the data into equidepth bins of depth 3.

  Bin 1: 13, 15, 16 Bin 2: 16, 19, 20 Bin 3: 20, 21, 22
  Bin 4: 22, 25, 25 Bin 5: 25, 25, 30 Bin 6: 33, 33, 35
  Bin 7: 35, 35, 35 Bin 8: 36, 40, 45 Bin 9: 46, 52, 70

- **Step 3:** Calculate the arithmetic mean of each bin.
- **Step 4:** Replace each of the values in each bin by the arithmetic mean calculated for the bin.

  Bin 1: 142/3, 142/3, 142/3 Bin 2: 181/3, 181/3, 181/3 Bin 3: 21, 21, 21
  Bin 4: 24, 24, 24 Bin 5: 262/3, 262/3, 262/3 Bin 6: 332/3, 332/3, 332/3
  Bin 7: 35, 35, 35 Bin 8: 401/3, 401/3, 401/3 Bin 9: 56, 56, 56

This method smooths a sorted data value by consulting to its "neighborhood". It performs *local* smoothing.

6. Use the methods below to *normalize* the following group of data:

$$200, 300, 400, 600, 1000$$

(a) min-max normalization by setting $min = 0$ and $max = 1$

(b) z-score normalization

(c) z-score normalization using the mean absolute deviation instead of standard deviation

(d) normalization by decimal scaling

(a) *min-max normalization* by setting $min = 0$ and $max = 1$ get the new value by computing

$$v_i' = \frac{v_i - 200}{1000 - 200}(1 - 0) + 0.$$

The normalized data are:

$$0, 0.125, 0.25, 0.5, 1$$

(b) In *z-score normalization*, a value v_i of A is normalized to v_i' by computing

$$v_i' = \frac{v_i - \bar{A}}{\sigma_A},$$

where

$$\bar{A} = \frac{1}{5}(200 + 300 + 400 + 600 + 1000) = 500,$$

$$\sigma_A = \sqrt{\frac{1}{5}(200^2 + 300^2 + ... + 1000^2) - \bar{A}^2} = 282.8.$$

The normalized data are:

$$-1.06, -0.707, -0.354, 0.354, 1.77$$

(c) *z-score normalization* using the *mean absolute deviation* instead of standard deviation replaces σ_A with s_A, where

$$s_A = \frac{1}{5}(|200 - 500| + |300 - 500| + ... + |1000 - 500|) = 240$$

The normalized data are:

$$-1.25, -0.833, -0.417, 0.417, 2.08$$

(d) The smallest integer j such that $Max(|\frac{v_i}{10^j}|) < 1$ is 3. After *normalization by decimal scaling*, the data become:

$$0.2, 0.3, 0.4, 0.6, 1.0$$

3. Exercise 2.2 gave the following data (in increasing order) for the attribute *age*: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

(a) Use min-max normalization to transform the value 35 for *age* onto the range $[0.0, 1.0]$.

(b) Use z-score normalization to transform the value 35 for *age*, where the standard deviation of *age* is 12.94 years.

(c) Use normalization by decimal scaling to transform the value 35 for *age*.

(d) Comment on which method you would prefer to use for the given data, giving reasons as to why.

(a) Use min-max normalization to transform the value 35 for *age* onto the range $[0.0, 1.0]$.

Using the corresponding equation with $min_A = 13$, $max_A = 70$, $new\_min_A = 0$, $new\_max_A = 1.0$, then $v = 35$ is transformed to $v' = 0.39$.

(b) Use z-score normalization to transform the value 35 for *age*, where the standard deviation of *age* is 12.94 years.

Using the corresponding equation where $A = 809/27 = 29.96$ and $\sigma_A = 12.94$, then $v = 35$ is transformed to $v' = 0.39$.

(c) Use normalization by decimal scaling to transform the value 35 for *age*.

Using the corresponding equation where $j = 2$, $v = 35$ is transformed to $v' = 0.35$.

. Suppose a hospital tested the age and body fat data for 18 randomly selected adults with the following result

age	23	23	27	27	39	41	47	49	50
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
age	52	54	54	56	57	58	58	60	61
%fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

(a) Normalize the two attributes based on *z-score normalization.*

(b) Calculate the *correlation coefficient* (Pearson's product moment coefficient). Are these two attributes positively or negatively correlated? Compute their covariance.

(a) Normalize the two variables based on *z-score normalization.*

age	23	23	27	27	39	41	47	49	50
z-age	-1.83	-1.83	-1.51	-1.51	-0.58	-0.42	0.04	0.20	0.28
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
z-%fat	-2.14	-0.25	-2.33	-1.22	0.29	-0.32	-0.15	-0.18	0.27
age	52	54	54	56	57	58	58	60	61
z-age	0.43	0.59	0.59	0.74	0.82	0.90	0.90	1.06	1.13
%fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7
z-%fat	0.65	1.53	0.0	0.51	0.16	0.59	0.46	1.38	0.77

(b) Calculate the *correlation coefficient* (Pearson's product moment coefficient). Are these two variables positively or negatively correlated?

The *correlation coefficient* is 0.82. The variables are positively correlated.

9. Suppose a group of 12 *sales price* records has been sorted as follows:

$$5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215.$$

Partition them into three bins by each of the following methods.

(a) equal-frequency (equidepth) partitioning

(b) equal-width partitioning

(c) clustering

Answer:

(a) equal-frequency (equidepth) partitioning

Partition the data into equidepth bins of depth 4:

Bin 1: 1: 5, 10, 11, 13 Bin 2: 15, 35, 50, 55 Bin 3: 72, 92, 204, 215

(b) equal-width partitioning

Partitioning the data into 3 equi-width bins will require the width to be $(215 - 5)/3 = 70$. We get:

Bin 1: 5, 10, 11, 13, 15, 35, 50, 55, 72 Bin 2: 92 Bin 3: 204, 215

(c) clustering

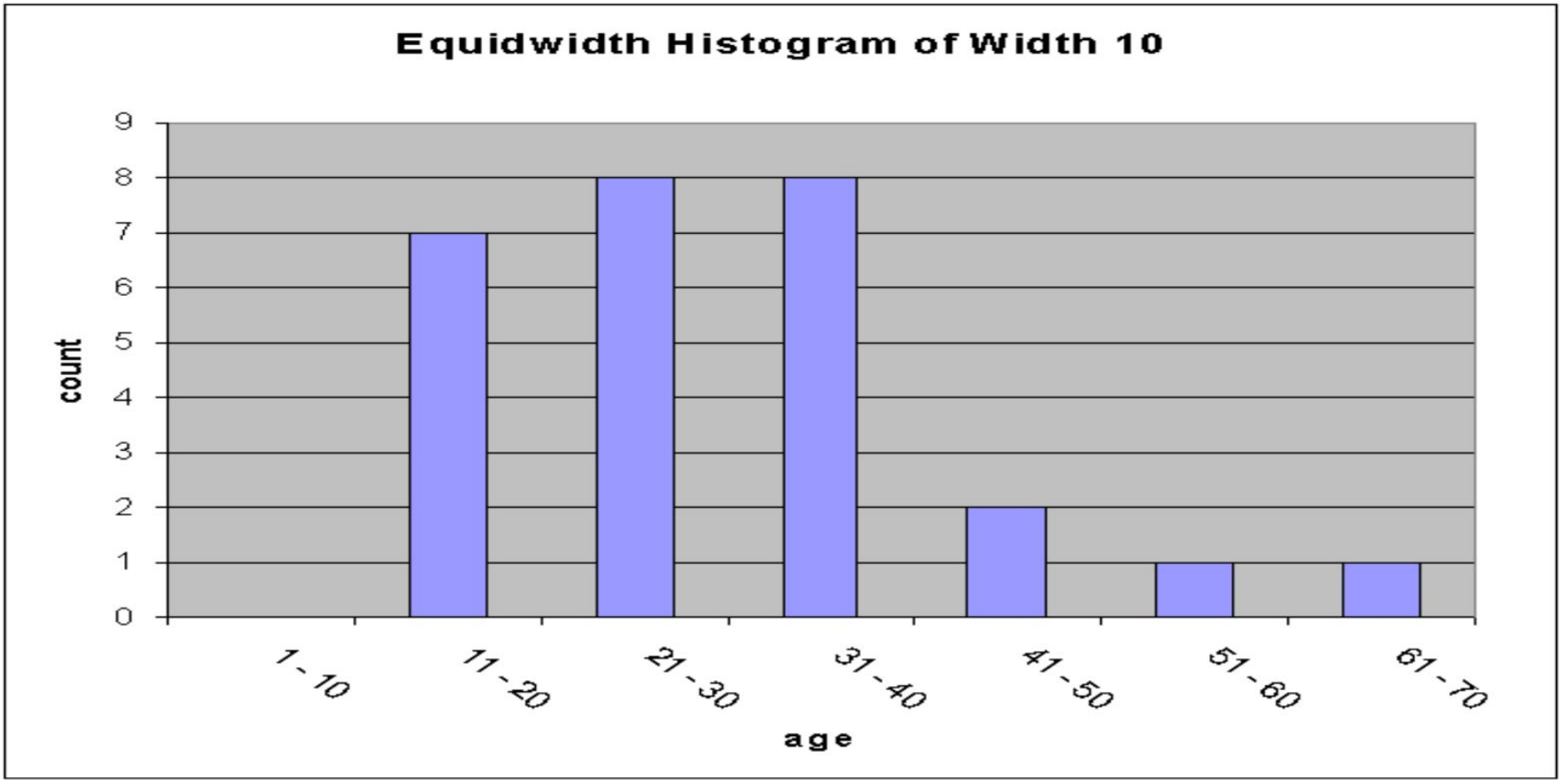
Using K-means clustering to partition the data into three bins we get:

Bin 1: 5, 10, 11, 13, 15, 35 Bin 2: 50, 55, 72, 92 Bin 3: 204, 215

3. Exercise 2.2 gave the following data (in increasing order) for the attribute *age*: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

(a) Plot an equal-width histogram of width 10.

(b) Sketch examples of each of the following sampling techniques: SRSWOR, SRSWR, cluster sampling, stratified sampling. Use samples of size 5 and the strata "youth", "middle-aged", and "senior".



Tuples

T1	13
T2	15
T3	16
T4	16
T5	19
T6	20
T7	20
T8	21
T9	22

T10	22
T11	25
T12	25
T13	25
T14	25
T15	30
T16	33
T17	33
T18	33

T19	33
T20	35
T21	35
T22	36
T23	40
T24	45
T25	46
T26	52
T27	70

SRSWOR vs. SRSWR

SRSWOR	(n = 5)
T4	16
T6	20
T10	22
T11	25
T26	32

SRSWR	(n = 5)
T7	20
T7	20
T20	35
T21	35
T25	46

Clustering sampling: Initial clusters

T1	13
T2	15
T3	16
T4	16
T5	19

T6	20
T7	20
T8	21
T9	22
T10	22

T11	25
T12	25
T13	25
T14	25
T15	30

T16	33
T17	33
T18	33
T19	33
T20	35

T21	35
T22	36
T23	40
T24	45
T25	46

T26	52
T27	70

Cluster sampling (m = 2)

T6	20
T7	20
T8	21
T9	22
T10	22

T21	35
T22	36
T23	40
T24	45
T25	46

Stratified Sampling

T1	13	young
T2	15	young
T3	16	young
T4	16	young
T5	19	young
T6	20	young
T7	20	young
T8	21	young
T9	22	young
T10	22	young
T11	25	young
T12	25	young
T13	25	young
T14	25	young
T15	30	middle age
T16	33	middle age
T17	33	middle age
T18	33	middle age
T19	33	middle age
T20	35	middle age
T21	35	middle age
T22	36	middle age
T23	40	middle age
T24	45	middle age
T25	46	middle age
T26	52	middle age
T27	70	senior

Stratified Sampling (according to age)

T4	16	young
T12	25	young
T17	33	middle age
T25	46	middle age
T27	70	senior

Concept Hierarchy Generation

- **Concept hierarchy** organizes concepts (i.e., attribute values) hierarchically and is usually associated with each dimension in a data warehouse
- Concept hierarchies facilitate drilling and rolling in data warehouses to view data in multiple granularity
- Concept hierarchy formation: Recursively reduce the data by collecting and replacing low level concepts (such as numeric values for *age*) by higher level concepts (such as *youth, adult*, or *senior*)
- Concept hierarchies can be explicitly specified by domain experts and/or data warehouse designers
- Concept hierarchy can be automatically formed for both numeric and nominal data. For numeric data, use discretization methods shown.

Concept Hierarchy Generation for Nominal Data

- Specification of a partial/total ordering of attributes explicitly at the schema level by users or experts
  - *street* < *city* < *state* < *country*
- Specification of a hierarchy for a set of values by explicit data grouping
  - {Urbana, Champaign, Chicago} < Illinois
- Specification of only a partial set of attributes
  - E.g., only *street* < *city*, not others
- Automatic generation of hierarchies (or attribute levels) by the analysis of the number of distinct values
  - E.g., for a set of attributes: {*street, city, state, country*}

Automatic Concept Hierarchy Generation

- Some hierarchies can be automatically generated based on the analysis of the number of distinct values per attribute in the data set
  - The attribute with the most distinct values is placed at the lowest level of the hierarchy
  - Exceptions, e.g., weekday, month, quarter, year