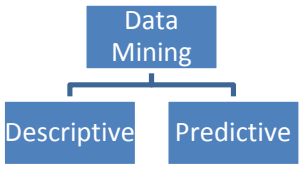
Unit 4

Data mining can be classified into two categories

1) **Descriptive mining**: describes concepts or task-relevant data sets in concise, summarative, informative, discriminative forms
2) **Predictive mining**: Based on data and analysis, constructs models for the database, and predicts the trend and properties of unknown data



CONCEPT DESCRIPTION:

Characterization: Provides a brief and clear summarization of the given collection of data.

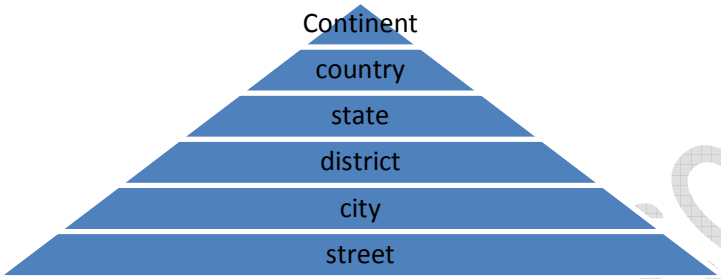
Comparison: Provides descriptions comparing two or more collections of data.

Difference between Concept Descriptions and OLAP

Functionality	Concept description	OLAP
Complex data types and aggregation	Concept Description can handle complex data types including numeric ,non numeric, spatial, text and can apply aggregation on them if necessary.	restricted to a small number of dimension and measure types. Many dimensions confine to non numeric data. And count (), Sum (), average () can only be applied to numeric data.
User-control versus automation	It is generally automated process	OLAP is user-controlled process. S;/ince user has to decide when to roll up, when to drill down etc..

DATA GENERALIZATION AND SUMMARIZATION BASED CHARACTERIZATION

Data generalization: It is a process which abstracts a large set of task-relevant data in a database from low conceptual levels to higher ones. For example the sales according to locations can be generalized as below.



There are mainly two approaches for data generalization:

1) Data cube approach (OLAP approach)
2) Attribute-oriented induction approach

ATTRIBUTE ORIENTED INDUCTION:

The attribute-oriented induction (AOI) approach to concept description was first proposed in 1989.The general idea behind attribute oriented induction is as follows.

STEP1: DATA FOCUSING:

Collect the task-relevant data by using data mining query on huge set of data. Selecting the relevant set of data not only makes mining more efficient, but also derives more meaningful results than mining the entire database.

For Example:

use Big University DB
mine characteristics as "Science Students"
in relevance to name, gender, major, birth place, birth date, residence,
phone gpa
from student
where status in "graduate"

The above query collect task relevant data .i.e data related to graduate students which results in following table.

Name	Gender	Major	Birth-Place	Birth_date	Residence	Phone #	GPA
Jim Woodman	M	CS	Vancouver,BC ,Canada	8-12-76	3511 Main St., Richmond	687-4598	3.67
Scott Lachance	M	CS	Montreal, Que, Canada	28-7-75	345 1st Ave., Richmond	253-9106	3.70
Laura Lee …	F …	Physics …	Seattle, WA, USA …	25-8-70 …	125 Austin Ave., Burnaby …	420-5232 …	3.83 …

STEP 2: DATA GENERALIZATION:

Depending on situation data generalization is performed by using following methods.

1. Attribute removal: An attribute is removed if there is a large set of distinct values for that attribute but (1) There is no generalization operator on it, or (2) Attribute's higher level concepts are expressed in terms of other attributes.
For Example attribute **NAME** can be removed from the table because it has a large set of distinct values. And an attribute **CITY** can be removed if it is already expressed in **LOCATION** attribute.

2**. Attribute generalization:** If there is a large set of distinct values for an attribute in the initial working relation, and there exists a set of generalization operators on the attribute, then a generalization operator should be selected and applied to the attribute. . For Example consider attribute **marks of students** (which is a continuous valued attribute) which can be generalized in terms of grades.

Some terms under attribute generalization:-

Attribute generalization control: The process of controlling the level of generalization for an attribute is known as attribute generalization control.

Attribute generalization threshold control: Setting one generalization threshold for all of the attributes, or sets one threshold for each attribute.

Generalized relation threshold control: unlike setting threshold for an attribute, Generalized threshold control sets a threshold for the generalized relation.

Attribute Name	**Action**	**Reason**
Name	Remove	Large Domain and cannot be generalized
Gender	Nothing	Very small domain
Major	Generalized	Generalized to arts, science, Engineering, business
Birth_place	Generalized (Removed and replaced with another attribute)	If generalized operator exists then generalize to country level and replace birth_place with Birth Country (should no cross the generalized relation threshold)
Birth_date	--do--	Birth_date (30-05-1987) > age(25) > age range(25…30)
Residence	---do--	Number > street > residence city > state > country
Phone	Removed	Large Domain and cannot be generalized
GPA	Generalized	{3.75 – 4.0 ,3.5 – 3.75 ….} > { Excellent , Very good…}

STEP 3: DATA AGGREGATION

It is done by merging identical, generalized tuples and accumulating their respective counts.

Gender	Major	Birth_Country	Age_range	Residence_city	GPA	Count
M	Science	Canada	20-25	Richmond	Very-good	16
F	Science	Foreign	25-30	Burnaby	Excellent	22
…	…	…	…	…	…	…

STEP 4: PRESENTATION OF THE GENERALIZED RELATION

Generalized relation (table)

A generalized relation for the sales in 2004.

location	*item*	*sales (in million dollars)*	count *(in thousands)*
Asia	TV	15	300
Europe	TV	12	250
North_America	TV	28	450
Asia	computer	120	1000
Europe	computer	150	1200
North_America	computer	200	1800

a) CROSS-TABULATION

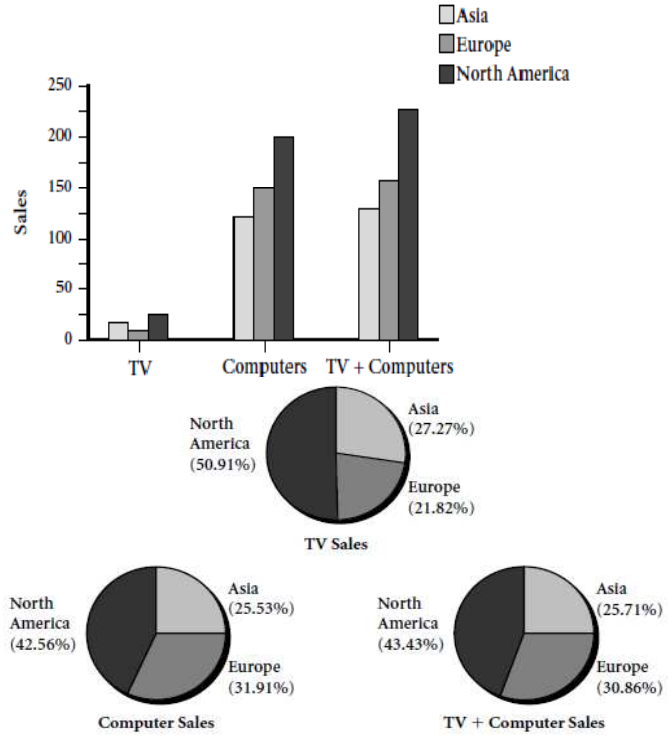
Descriptions can also be visualized in the form of cross-tabulations, or crosstabs. In a two-dimensional crosstab, each row represents a value from an attribute, and each column represents a value from another attribute.

A crosstab for the sales in 2004.

	item					
	TV		*computer*		*both_items*	
location	sales	count	sales	count	sales	count
Asia	15	300	120	1000	135	1300
Europe	12	250	150	1200	162	1450
North_America	28	450	200	1800	228	2250
all_regions	45	1000	470	4000	525	5000

b) BAR CHART AND PIE CHART

The sales data of the crosstab shown in above table can be transformed into the bar chart as well as pie chart.



Pie chart representation of the sales in 2004.

c) REPRESENTATION IN TERM OF LOGICAL RULES:

Generalized relation may also be represented in the form of logic rules. These rules are also known as quantitative rule (since they represent some quantity).
When a quantitative rule represents some generalized information regarding target class of data then it is known as quantitative characteristic rule.

T-weight is an interestingness measure that describes the typicality of each disjunct in the rule. The t-weight is represented as

$$t\_weight = \text{count}(q_a)/\Sigma_{i=1}^{n}\text{count}(q_a),$$

And the quantitative characteristic rule should be expressed in the form

$$\forall X,\ target\_class(X) \Rightarrow condition_1(X)[t : w_1] \vee \cdots \vee condition_m(X)[t : w_m].$$

Example of Quantitative characteristic rule:

$$\forall X,\ item(X) = \text{“}computer\text{”} \Rightarrow$$
$$(location(X) = \text{“}Asia\text{”})\ [t : 25.00\%] \vee (location(X) = \text{“}Europe\text{”})\ [t : 30.00\%] \vee$$
$$(location(X) = \text{“}North\_America\text{”})\ [t : 45,00\%]$$

Here t:25.00% is **number of computers sold in Asia (1000) / total computer sold (4000)**

Algorithm: Attribute oriented induction. Mining generalized characteristics in a relational database given a user's data mining request.

Input:

DB*,* a relational database;
DMQuery*,* a data mining query;
a list, a list of attributes (containing attributes, *ai*);
***Gen*(*ai*),** a set of concept hierarchies or generalization operators on attributes, *ai*;
***a gen thresh*(*ai*)**, attribute generalization thresholds for each *ai*.

Output: *P*, a *Prime generalized relation*.
Method:

1. *W* ⟵ get task relevant data (*DMQuery, DB*); // Let *W*, the working relation, hold the task-relevant data.

2. Prepare for generalization **(*W*);** // This is implemented as follows.
(a) Scan ***W*** and collect the distinct values for each attribute**, *ai***. (Note: If ***W*** is very large, this may be done by examining a sample of ***W***.)

(b) For each attribute ***ai*,** determine whether ***ai*** should be removed, and if not, compute its minimum desired level ***Li*** based on its given or default attribute threshold, and determine the mapping pairs **(*v*, *v*'),** where *v* is a distinct value of ***ai*** in ***W***, and ***v*'** is its corresponding generalized value at level *Li*.

3. *P* ⟵ generalization (***W***),
The *Prime generalized relation*, ***P***, is derived by replacing each value ***v*** in ***W*** by its corresponding ***v*'** in the mapping while accumulating count and computing any other aggregate values.

This step can be implemented efficiently using either of the two following variations:

(a) For each generalized tuple, insert the tuple into a sorted prime relation ***P*** by a binary search: if the tuple is already in ***P***, simply increase its count and other aggregate values accordingly; otherwise, insert it into *P*.

(b) Since in most cases the number of distinct values at the prime relation level is small, the prime relation can be coded as an ***m*-dimensional array** where ***m*** is the number of attributes in ***P*,** and each dimension contains the corresponding generalized attribute values. Each array element holds the corresponding count and other aggregation values, if any. The insertion of a generalized tuple is performed by measure aggregation in the corresponding array element.

ANALYTICAL CHARACTERIZATION: ANALYSIS OF ATTRIBUTE RELEVANCE

Why perform attribute relevance analysis?

It is very important to know which attribute is relevant for analysis and which attribute is not. Methods should be introduced to perform attribute (or dimension)relevance Analysis in order to filter out statistically irrelevant or weakly relevant attributes, and retain or even rank the most relevant attributes for the descriptive mining task at hand.
Class characterization that includes the analysis of attribute/dimension relevance is called **analytical characterization**. Class comparison that includes such analysis is called analytical comparison. An attribute or dimension is considered highly relevant with respect to a given class if it is likely that the values of the attribute or dimension may be used to distinguish the class from others.

Methods for attribute relevance

1) Information gain (ID3)
2) Gain ratio (C4.5)
3) Gini index
4) Contingency table statistics
5) Uncertainty coefficient

Information gain (ID3) is mainly used method for attribute relevance analysis.
It is done as follows

STEP 1. DATA COLLECTION

Collect data for both the target class(graduate students) and the contrasting class(under graduate) by query processing. The target class is the class to be characterized, whereas the contrasting class is the set of comparable data that are not in the target class.

Fig: Target class data

name	*gender*	*major*	*birth_place*	*birth_date*	*residence*	*phone#*	*gpa*
Jim Woodman	M	CS	Vancouver, BC, Canada	8-12-76	3511 Main St., Richmond	687-4598	3.67
Scott Lachance	M	CS	Montreal, Que, Canada	28-7-75	345 1st Ave., Vancouver	253-9106	3.70
Laura Lee	F	Physics	Seattle, WA, USA	25-8-70	125 Austin Ave., Burnaby	420-5232	3.83
...	...	...	...	...	...	...	...

Fig: contrasting class data

name	gender	major	birth_place	birth_date	residence	phone#	gpa
Bob Schumann	M	Chemistry	Calgary, Alt, Canada	10-1-78	2642 Halifax St., Burnaby	294-4291	2.96
Amy Eau	F	Biology	Golden, BC, Canada	30-3-76	463 Sunset Cres., Vancouver	681-5417	3.52
...	...	...	...	...	...	...	...

STEP 2: PRELIMINARY RELEVANCE ANALYSIS BY APPLYING CONSERVATIVE AOI

In this step preliminary analysis is done by attribute oriented induction and unnecessary attribute such as name, phone no. are removed. Attributes are also generalized by keeping low generalization threshold value.

gender	major	birth_country	age_range	gpa	count
M	Science	Canada	20-25	Very_good	16
F	Science	Foreign	25-30	Excellent	22
M	Engineering	Foreign	25-30	Excellent	18
F	Science	Foreign	25-30	Excellent	25
M	Science	Canada	20-25	Excellent	21
F	Engineering	Canada	20-25	Excellent	18

Total graduate students:120

gender	major	birth_country	age_range	gpa	count
M	Science	Foreign	<20	Very_good	18
F	Business	Canada	<20	Fair	20
M	Business	Canada	<20	Fair	22
F	Science	Canada	20-25	Fair	24
M	Engineering	Foreign	20-25	Very_good	22
F	Engineering	Canada	<20	Excellent	24

Total undergraduate students 130

STEP 3: REMOVE IRRELEVANT OR WEEKLY RELEVANT ATTRIBUTES USING THE RELEVANCE MEASURE.

Here we are using information gain as the relevant measure, which is performed as follows.

1) Calculate expected information required to classify an arbitrary tuple is given by

$$I(s_1,s_2,\ldots,s_m) = -\sum_{i=1}^{m} \frac{s_i}{s} log_2 \frac{s_i}{s}$$

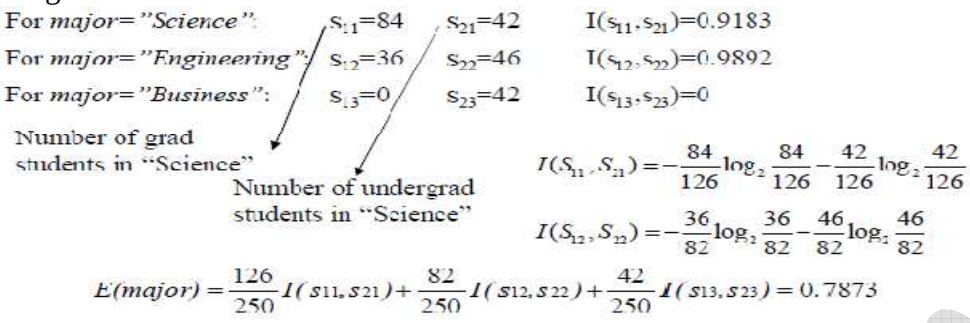
Where' S' is total number of samples si are the samples of class 'ci'. By substituting the values we get.

$$I(s_1,s_2) = I(120,130) = -\frac{120}{250} log_2 \frac{120}{250} - \frac{130}{250} log_2 \frac{130}{250} = 0.9988$$

2) Calculate entropy of each attribute: e.g. major

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j}+\ldots+s_{mj}}{s} I(s_{1j},\ldots,s_{mj})$$

By substituting the values.



3) Information gain obtained by partitioning on attribute A

$$Gain(A) = I(s_1, s_2, \ldots, s_m) - E(A)$$

By substituting the values:

$$Gain(major) = I(s_1, s_2) - E(major) = 0.2115$$

In the similar manner the information gain of rest of the attributes are

Gain (gender) = 0.0003
Gain (birth_country) = 0.0407
Gain (major) = 0.2115
Gain (gpa) = 0.4490
Gain (age_range) = 0.5971

Now, remove irrelevant/weakly relevant attributes from candidate which do not satisfy the threshold value 0.1 drop gender, birth_country and remove contrasting class candidate relation. This data is known as initial target class working relation. This is shown below.

major	age_range	gpa	count
Science	20-25	Very_good	16
Science	25-30	Excellent	47
Science	20-25	Excellent	21
Engineering	20-25	Excellent	18
Engineering	25-30	Excellent	18

Initial target class working relation W_0: Graduate students

STEP 4: GENERATE THE CONCEPT DESCRIPTION USING ATTRIBUTE ORIENTED INDUCTION.

Perform AOI using less conservative set of AOI threshold.

MINING CLASS COMPARISONS: DISCRIMINATING BETWEEN DIFFERENT CLASSES

Class discrimination or comparison (hereafter referred to as class comparison) mines descriptions that distinguish a target class from its contrasting classes. Notice that the target and contrasting classes must be *comparable* in the sense that they share similar dimensions and attributes. For example, the three classes, *person, address,* and *item,* are not comparable. Attribute generalization process described for class characterization can be modified so that the generalization is performed *synchronously* among all the classes compared. This allows the attributes in all of the classes to be generalized to the *same* levels of abstraction.

1. **DATA COLLECTION:** The set of relevant data in the database is collected by query processing and is partitioned respectively into a *target class* and one or a set of *contrasting class(es).*

   use Big_University_DB
   mine comparison as "grad_vs_undergrad_students"
   in relevance to *name, gender, major, birth_place, birth_date, residence, phone#, gpa*
   for "graduate_students"
   where status in "graduate"
   versus "undergraduate_students"
   where status in "undergraduate"
   analyze count%
   from student

   Initial target class

name	gender	major	birth_place	birth_date	residence	phone#	gpa
Jim Woodman	M	CS	Vancouver, BC, Canada	8-12-76	3511 Main St., Richmond	687-4598	3.67
Scott Lachance	M	CS	Montreal, Que, Canada	28-7-75	345 1st Ave., Vancouver	253-9106	3.70
Laura Lee	F	Physics	Seattle, WA, USA	25-8-70	125 Austin Ave., Burnaby	420-5232	3.83
...	...	...	...	...	...	...	...

   Initial Contrasting Class

name	gender	major	birth_place	birth_date	residence	phone#	gpa
Bob Schumann	M	Chemistry	Calgary, Alt, Canada	10-1-78	2642 Halifax St., Burnaby	294-4291	2.96
Amy Eau	F	Biology	Golden, BC, Canada	30-3-76	463 Sunset Cres., Vancouver	681-5417	3.52
...	...	...	...	...	...	...	...

2. **DIMENSION RELEVANCE ANALYSIS**: If there are many dimensions, then dimension relevance analysis should be performed on these classes to select only the highly relevant dimensions for further analysis. Correlation or **entropy-based measures** can be used for this step. Irrelevant or weakly relevant dimensions, such as *name, gender, birth place, residence,* and *phone#,* are removed from the resulting classes. Only the highly relevant attributes are included in the subsequent analysis.

3. **SYNCHRONOUS GENERALIZATION**: Generalization is performed on the target class to the level controlled by a user- or expert-specified dimension threshold, which results in a prime target class relation. The concepts in the contrasting class (es) are generalized to the same level as those in the prime target class relation, forming the prime contrasting class(es) relation.

Prime generalized relation for the *contrasting class* (undergraduate students)

major	age_range	gpa	count%
Science	16...20	fair	5.53%
Science	16...20	good	4.53%
...	...	...	...
Science	26...30	good	2.32%
...	...	...	...
Business	over_30	excellent	0.68%

Prime generalized relation for the *target class* (graduate students)

major	age_range	gpa	count%
Science	21...25	good	5.53%
Science	26...30	good	5.02%
Science	over_30	very_good	5.86%
...	...	...	...
Business	over_30	excellent	4.68%

4. **Presentation of the derived comparison**: The resulting class comparison description can be visualized in the form of tables, graphs, and rules. This presentation usually includes a "contrasting" measure such as count% (percentage count) that reflects the comparison between the target and contrasting classes. The user can adjust the comparison description by applying drill-down, roll-up, and other OLAP operations to the target and contrasting classes, as desired.

A quantitative discriminant rule for the target class of a given comparison description is written in the form.

$$\forall X,\ target\_class(X) \Leftarrow condition(X)\ \ [d\!:\!d\_weight],$$

The d-weight for ***qa*** is the ratio of the number of tuples from the initial target class working relation that are covered by ***qa*** to the total number of tuples in both the initial target class and contrasting class working relations that are covered by ***qa***. Formally, the d-weight of ***qa*** for the class *Cj* is defined as,

$$d\_weight = \text{count}(q_a \in C_j) / \Sigma_{i=1}^{m} \text{count}(q_a \in C_i),$$

For Example,

Count distribution between graduate and undergraduate students for a generalized tuple.

status	major	age_range	gpa	count
graduate	Science	21...25	good	90
undergraduate	Science	21...25	good	210

$$\forall X,\ Status(X) = \text{“}graduate\_student\text{”} \Leftarrow major(X) = \text{“}Science\text{”} \wedge age\_range(X) = \text{“}21...25\text{”} \wedge gpa(X) = \text{“}good\text{”}[d : 30\%].$$

Note: d-weight=**Graduate (90) / (Graduate (90) +undergraduate (210))**

CLASS DESCRIPTION: PRESENTATION OF BOTH CHARACTERIZATION AND COMPARISON

QUANTITATIVE DESCRIPTION RULE: A quantitative characteristic rule and a quantitative discriminant rule for the same class can be combined to form a *quantitative description rule* for the class, which displays the t-weights *and* d-weights associated with the corresponding characteristic and discriminant rules.

The *quantitative description rule is expressed in te form,*

$$\forall X,\ target\_class(X) \Leftrightarrow condition_1(X)[t : w_1, d : w'_1]$$
$$\theta \cdots \theta\ condition_m(X)[t : w_m, d : w'_m],$$

A crosstab for the total number (count) of TVs and computers sold in thousands in 2004.

	item		
location	TV	computer	both_items
Europe	80	240	320
North_America	120	560	680
both_regions	200	800	1000

The above table can be expressed in with T-weight and D-weight as follows,

	item								
	TV			computer			both_items		
location	count	t-weight	d-weight	count	t-weight	d-weight	count	t-weight	d-weight
Europe	80	25%	40%	240	75%	30%	320	100%	32%
North_America	120	17.65%	60%	560	82.35%	70%	680	100%	68%
both_regions	200	20%	100%	800	80%	100%	1000	100%	100%

The quantitative description rule for the target class, *Europe*, is

$$\forall X,\ location(X) = \text{“}Europe\text{”} \Leftrightarrow$$
$$(item(X) = \text{“}TV\text{”})\ [t : 25\%, d : 40\%]\ \theta\ (item(X) = \text{“}computer\text{”})$$
$$[t : 75\%, d : 30\%].$$

Note: **t-weight of TV in Europe** = (sales of tv in Europe)/(sales of tv in europe+sales of computer in europe)

d-weight of TV in Europe = (sales of tv in Europe)/(sales of tv in Europe +Sales of TV in north America)

PLEASE WRITE T-WEIGTH AND D-WEIGHT FORMULA WITH THIS CONTENT.

MINING DESCRIPTIVE STATISTICAL MEASURES IN LARGE DATABASES

Relational database systems provide built-in aggregate functions: count(), sum(), avg(), max(), and min(). But, for doing data mining we need more functions to describe the central tendency and dispersion of data.

Measures of central tendency include mean, median, mode, and midrange, while measures of data dispersion include quartiles, outliers, variance, and other statistical measures.

MEASURING THE CENTRAL TENDENCY

The most common and most effective numerical measure of the "center" of a set of data is the (arithmetic) mean. Let x1, x2…… xn be a set of n values or observations. The mean of this set of values is

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

Sometimes, each value xi in a set may be associated with a weight wi, for i = 1…n. The weights reflect the significance, importance, or occurrence frequency attached to their respective values. In this case, we can compute the weighted arithmetic mean as,

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

For skewed data, a better measure of center of data is the median, M. Suppose that the values forming a given set of data are in numerical order. The median is the middle value of the ordered set if the number of values n is an odd number; otherwise (i.e.if n is even), it is the average of the middle two values. For example, for grouped data, the median, obtained by interpolation, is given by

$$median = L_1 + (\frac{n/2 + (\sum f)_l}{f_{median}})c.$$

The mode for a set of data is the value that occurs most frequently in the set. Data sets with one, two, or three modes are respectively called unimodal, bimodal, and trimodal. If a data set has more than three modes, it is multimodal. For unimodal frequency curves that are moderately skewed (asymmetrical), we have the following empirical relation

$$mean \Leftrightarrow mode = 3 \times (mean \Leftrightarrow median)..$$

The midrange, that is, the average of the largest and smallest values in a data set, can be used to measure the central tendency of the set of data.

MEASURING THE DISPERSION OF DATA

The most common measures of data dispersion are the five-number summary (based on quartiles), the interquartile range, and standard deviation. The plotting of boxplots (which show outlier values) also serves as a useful graphical method.

The first quartile, denoted by Q1, is the 25-th percentile; and the third quartile, denoted by Q3, is the 75-th percentile.

INTERQUARTILE RANGE: The distance between the first and third quartiles is a simple measure of spread that gives the range covered by the middle half of the data. This distance is called the interquartile range (IQR), and is defined as,

$$IQR = Q_3 \Leftrightarrow Q_1.$$

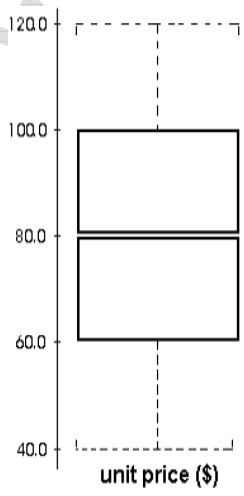
One common rule of thumb for identifying suspected outliers is to single out values falling at least 1.5 * IQR above the third quartile or below the first quartile.

The five number summary of a distribution consists of the median M, the quartiles Q1 and Q3, and the smallest and largest individual observations, written in the order

Minimum, Q1, M, Q3, Maximum

A popularly used visual representation of a distribution is the boxplot. In a boxplot:
1. The ends of the box are at the quartiles, so that the box length is the interquartile range, IQR.
2. The median is marked by a line within the box.
3. Two lines (called whiskers) outside the box extend to the smallest (Minimum) and largest (Maximum) observations.

Figure



VARIANCE AND STANDARD DEVIATION

The variance of n observations x1, x2…. xn is

$$s^2 = \frac{1}{n-1}\left[\sum x_i^2 - \frac{1}{n}(\sum x_i)^2\right]$$

Standard deviation: the square root of the variance