Getting to Know Your Data

- **Data Objects and Attribute Types**
- **Basic Statistical Descriptions of Data**
- **Measuring Data Similarity and Dissimilarity**

- **Data Objects and Attribute Types**
- Basic Statistical Descriptions of Data
- Measuring Data Similarity and Dissimilarity

Data-Related Issues for Successful Data Mining

Type of Data:

- Data sets differ in a number of ways.
- Type of data determines which techniques can be used to analyze the data.

Quality of Data:

- Data is often far from perfect.
- Improving data quality improves the quality of the resulting analysis.

Preprocessing Steps to Make Data More Suitable for Data Mining:

- Raw data must be processed in order to make it suitable for analysis.
  - Improve data quality,
  - Modify data so that it better fits a specified data mining technique.

Analyzing Data in Terms of its Relationships:

- find relationships among data objects and then perform remaining analysis using these relationships rather than data objects themselves.
- There are many similarity or distance measures, and the proper choice depends on the type of data and application.

What is Data?

- **Data sets** are made up of **data objects**.
- A **data object** represents an entity.
  - Also called ***sample, example, instance, data point, object, tuple.***
- Data objects are described by **attributes**.
- An **attribute** is a property or characteristic of a *data object.*
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as ***variable***, ***field***, ***characteristic***, or ***feature***
- A collection of attributes describe an object.
- **Attribute values** are numbers or symbols assigned to an attribute.

A Data Object

Attributes

Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- database rows ➔ data objects
- database columns ➔ attributes

Attributes

- **Attribute** (or dimensions, features, variables): a data field, representing a characteristic or feature of a data object.
  - E.g., customer _ID, name, address
- **Attribute values** are numbers or symbols assigned to an attribute

- Distinction between attributes and attribute values
  - Same attribute can be mapped to different attribute values
    - Example: height can be measured in feet or meters
  - Different attributes can be mapped to the same set of values
    - Example: Attribute values for ID and age are integers
    - But properties of attribute values can be different; ID has no limit but age has a maximum and minimum value

Attribute Types

Four main types of attributes

Nominal: Categorical (Qualitative)

- categories, states, or "names of things"
  - Hair color, marital status, occupation, ID numbers, zip codes
- An important nominal attribute: **Binary**
  - Nominal attribute with only 2 states (0 and 1)

Ordinal: Categorical (Qualitative)

- Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - Size = {small, medium, large}, grades, army rankings

Interval: Numeric (Quantitative)

- Measured on a scale of equal-sized units
- Values have order:
  - temperature in C˚ or F˚, calendar dates
- No true zero-point: ratios are not meaningful

Ratio: Numeric (Quantitative)

- Inherent zero-point: ratios are meaningful
  - temperature in Kelvin, length, counts, monetary quantities

Attribute Types

Four main types of attributes: Nominal Attributes

- The values of a **nominal attribute** are symbols or names of things.
  - Each value represents some kind of category, code, or state,
- Nominal attributes are also referred to as **categorical attributes**.
- The values of nominal attributes do not have any meaningful order.
- Example: The attribute ***marital_status*** can take on the values ***single***, ***married***, ***divorced***, and ***widowed***.

- Because **nominal attribute** values do not have any meaningful order about them and they are not quantitative.
  - It makes no sense to find the ***mean (average)*** value or ***median (middle)*** value for such an attribute.
  - However, we can find the attribute's most commonly occurring value (***mode***).

Attribute Types

Four main types of attributes: Nominal Attributes

- A **binary attribute** is a special ***nominal attribute*** with only two states: 0 or 1.

- A binary attribute is **symmetric** if both of its states are equally valuable and carry the same weight.
  - Example: the attribute ***gender*** having the states ***male*** and ***female***.

- A binary attribute is **asymmetric** if the outcomes of the states are not equally important.
  - Example: ***Positive*** and ***negative*** outcomes of a medical test for HIV.
  - By convention, we code the most important outcome, which is usually the rarest one, by 1 (e.g., HIV positive) and the other by 0 (e.g., HIV negative).

Attribute Types

Four main types of attributes: Ordinal Attributes

- An **ordinal attribute** is an attribute with possible values that have a meaningful order or ranking among them, but the magnitude between successive values is not known.
- Example: An ordinal attribute ***drink_size*** corresponds to the size of drinks available at a fast-food restaurant.
  - This attribute has three possible values: ***small***, medium, and ***large***.
  - The values have a meaningful sequence (which corresponds to increasing drink size); however, we cannot tell from the values how much bigger, say, a medium is than a large.
- The central tendency of an ordinal attribute can be represented by its ***mode*** and its ***median*** (middle value in an ordered sequence), but the ***mean*** cannot be defined.

Attribute Types

Four main types of attributes: Interval Attributes

- **Interval attributes** are measured on a ***scale of equal-size units***.
  - We can compare and quantify the difference between values of interval attributes.
- Example: A ***temperature*** attribute is an interval attribute.
  - We can quantify the difference between values. For example, a temperature of 20^{o}C is five degrees higher than a temperature of 15^{o}C.
  - Temperatures in Celsius do not have a **true zero-point**, that is, 0^{o}C does not indicate "no temperature."
  - Although we can compute the difference between temperature values, we cannot talk of one temperature value as being a multiple of another.
    - Without a true zero, we cannot say, for instance, that 10^{o}C is twice as warm as 5^{o}C . That is, we cannot speak of the values in terms of ratios.
- The central tendency of an interval attribute can be represented by its ***mode***, its ***median*** (middle value in an ordered sequence), and its ***mean***.

Attribute Types

Four main types of attributes: Ratio Attributes

- A **ratio attribute** is a numeric attribute with an *inherent zero-point*.
- Example: A ***number_of_words*** attribute is a ratio attribute.
  - If a measurement is ratio-scaled, we can speak of a value as being a multiple (or ratio) of another value.
- The central tendency of an ratio attribute can be represented by its ***mode***, its ***median*** (middle value in an ordered sequence), and its ***mean***.

Properties of Attribute Values

- The type of an attribute depends on which of the following properties it possesses:
  - Distinctness: = ≠
  - Order: < >
  - Addition: + -
  - Multiplication: * /

- Nominal attribute: **distinctness**
- Ordinal attribute: **distinctness & order**
- Interval attribute: **distinctness, order & addition**
- Ratio attribute: **all 4 properties**

Properties of Attribute Values

Attribute Type	Description	Examples
Nominal	The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male*, *female*}
Ordinal	The values of an ordinal attribute provide enough information to order objects. (<, >)	hardness of minerals, {*good*, *better*, *best*}, grades, street numbers
Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -)	calendar dates, temperature in Celsius or Fahrenheit
Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length,

Attribute Types
Categorical (Qualitative) and Numeric (Quantitative)

- **Nominal** and **Ordinal** attributes are collectively referred to as ***categorical or qualitative attributes***.
  - qualitative attributes, such as employee ID, lack most of the properties of numbers.
  - Even if they are represented by numbers, i.e. , integers, they should be treated more like symbols .
  - *Mean* of values does not have any meaning.

- **Interval** and **Ratio** are collectively referred to as ***quantitative or numeric attributes.***
  - Quantitative attributes are represented by numbers and have most of the properties of numbers .
  - Note that quantitative attributes can be integer-valued or continuous.
  - Numeric operations such as *mean*, *standard deviation* are meaningful

Discrete vs. Continuous Attributes

- **Discrete Attribute**
  - Has only a finite or countably infinite set of values
    - zip codes, profession, or the set of words in a collection of documents
  - Sometimes, represented as integer variables
  - Note: Binary attributes are a special case of discrete attributes
  - Binary attributes where only non-zero values are important are called **asymmetric binary attributes.**

- **Continuous Attribute**
  - Has real numbers as attribute values
    - temperature, height, or weight
  - Practically, real values can only be measured and represented using a finite number of digits
  - Continuous attributes are typically represented as floating-point variables

Types of data sets

- **Record**
  - Relational records
  - Data matrix, e.g., numerical matrix, crosstabs
  - Document data: text documents: term-frequency vector
  - Transaction data
- **Graph and network**
  - World Wide Web
  - Social or information networks
  - Molecular Structures
- **Ordered**
  - Video data: sequence of images
  - Temporal data: time-series
  - Sequential Data: transaction sequences
  - Genetic sequence data
- **Spatial, image and multimedia:**
  - Spatial data: maps
  - Image data:
  - Video data:

Record Data

- Data that consists of a collection of records, each of which consists of a fixed set of attributes

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Data Matrix

- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute
- Such data set can be represented by an m by n matrix, where there are m rows, one for each object, and n columns, one for each attribute.
- A data matrix is a variation of record data, but because it consists of numeric attributes, standard matrix operation can be applied to transform and manipulate the data.

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Document (Text) Data

- Each document becomes a ***term*** vector,
  - each term is a component (attribute) of the vector,
  - the value of each component is the number of times the corresponding term occurs in the document
- Convert text documents to record data by counting word frequencies (***document-term matrix***).

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Transaction Data

- ***Transaction data*** is a special type of record data, where
  - each record (***transaction***) involves a set of ***items***.
  - Example: The set of products purchased by a customer constitute a ***transaction***, while the individual products that were purchased are the ***items***.

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Transaction Data
Convert to Record Data

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Converted to record data

Requires less space

TID	Bread?	Coke?	Milk?	Diaper?	Beer?
1	1	1	1	0	0
2	1	0	0	0	1
3	0	1	1	1	1
4	1	0	1	1	1
5	0	1	1	1	0

Asymmetric attributes

Requires more space

- In real-world data, the table would contain hundreds or thousands of columns, depending on the number of items to be considered.
- The number of items bought in a transaction, say 5, is very small in comparison to the number of columns
- Most values in this matrix are “0”. Such a matrix is called sparse matrix.

- Data Objects and Attribute Types
- **Basic Statistical Descriptions of Data**
- Measuring Data Similarity and Dissimilarity

Basic Statistical Descriptions of Data

- **Basic statistical descriptions** can be used to identify properties of the data and highlight which data values should be treated as noise or outliers.

- For data preprocessing tasks, we want to learn about data characteristics regarding both **central tendency** and **dispersion of the data**.
- **Measures of central tendency** include **mean**, **median**, **mode**, and **midrange**.
- **Measures of data dispersion** include **quartiles**, **interquartile range (IQR)**, and **variance**.

- These descriptive statistics are of great help in understanding the distribution of the data.

Measuring Central Tendency: Mean

- The most common and most effective numerical measure of the "***center***" of a set of data is the **arithmetic mean**.

Arithmetic Mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

- Sometimes, each value x_i in a set may be associated with a weight w_i.
  - The weights reflect the significance and importance attached to their respective values.

Weighted Arithmetic Mean: $\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$

Measuring Central Tendency: Mean

- Although the mean is the single most useful quantity for describing a data set, it is not always the best way of measuring the center of the data.
  - A major problem with the mean is its sensitivity to extreme (outlier) values.
  - Even a small number of extreme values can corrupt the mean.
- To offset the effect caused by a small number of extreme values, we can instead use the **trimmed mean**,
- **Trimmed mean** can be obtained after chopping off values at the high and low extremes.

Measuring Central Tendency: Median

- Another measure of the center of data is the **median**.
- Suppose that a given data set of N distinct values is sorted in numerical order.
  - If N is odd, the **median** is the middle value of the ordered set;
  - If N is even, the **median** is the average of the middle two values.

- In probability and statistics, the **median** generally applies to numeric data; however, we may extend the concept to **ordinal data**.
  - Suppose that a given data set of N values for an attribute X is sorted in increasing order.
  - If N is odd, then the **median** is the middle value of the ordered set.
  - If N is even, then the **median** may not be not unique.
    - In this case, the median is the two middlemost values and any value in between.

Measuring Central Tendency: Mode

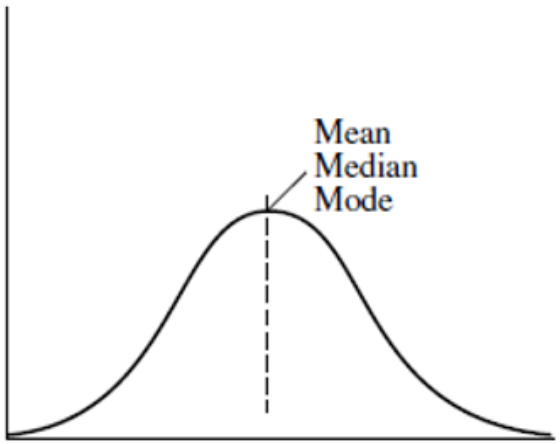
- Another measure of central tendency is the **mode**.
- The **mode** for a set of data is the value that occurs most frequently in the set.
  - It is possible for the greatest frequency to correspond to several different values, which results in more than one mode.
  - Data sets with one, two, or three modes: called ***unimodal***, ***bimodal***, and ***trimodal***.
  - At the other extreme, if each data value occurs only once, then there is no mode.

- *Central Tendency Measures for Numerical Attributes:* ***Mean, Median, Mode***

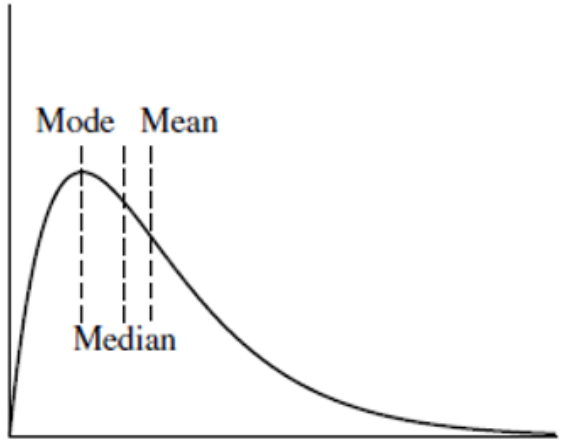
- *Central Tendency Measures for Categorical Attributes:* ***Mode*** *(Median?)*
  - *Central Tendency Measures for Nominal Attributes:* ***Mode***
  - *Central Tendency Measures for Ordinal Attributes:* ***Mode, Median***

Measuring Central Tendency - Mean, Median, Mode

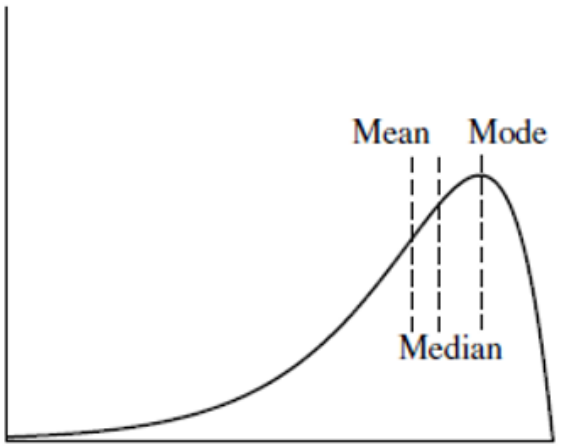
Median, ***mean*** and ***mode*** of *symmetric*, *positively* and *negatively skewed data*



symmetric data



positively skewed data



negatively skewed data

Measuring Central Tendency: Example

What are central tendency measures (mean, median, mode)for the following attributes?

attr1 = {2,4,4,6,8,24}

attr2 = {2,4,7,10,12}

attr3 = {xs,s,s,s,m,m,l}

Measuring Central Tendency: Example

What are central tendency measures (mean, median, mode)for the following attributes?

attr1 = {2,4,4,6,8,24}

mean = (2+4+4+6+8+24)/6 = 8	average of all values
median = (4+6)/2 = 5	avg. of two middle values
mode = 4	most frequent item

attr2 = {2,4,7,10,12}

mean = (2+4+7+10+12)/5 = 7	average of all values
median = 7	middle value
mode = any of them (no mode)	all of them has same freq.

attr3 = {xs,s,s,s,m,m,l}

mean is meaningless for categorical attributes.

median = s	middle value
mode = s	most frequent item

Measuring Dispersion of Data

- **The degree to which numerical data tend to spread is called the dispersion, or variance of the data.**

The most common ***measures of data dispersion:***

- **Range:** Difference between the largest and smallest values.
- **Interquartile Range (IQR):** range of middle 50%
  - **quartiles**: Q1 (25th percentile), Q3 (75th percentile) IQR=Q3-Q1
  - **five number summary**: Minimum, Q1, Median, Q3, Maximum
- **Variance and Standard Deviation:** (*sample: s, population: σ)*
  - **variance** of N observations:

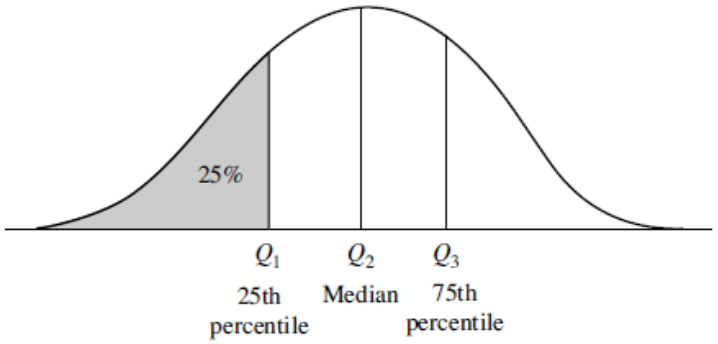
$$\sigma^2 = \frac{1}{n}\sum_{1}^{n}(x_i - \mu)^2 \qquad s^2 = \frac{1}{n-1}\sum_{1}^{n}(x_i - \mu)^2$$

where μ is the mean value of the observations

  - **standard deviation** σ *(s)* is the square root of variance σ^2 *(*s^2*)*

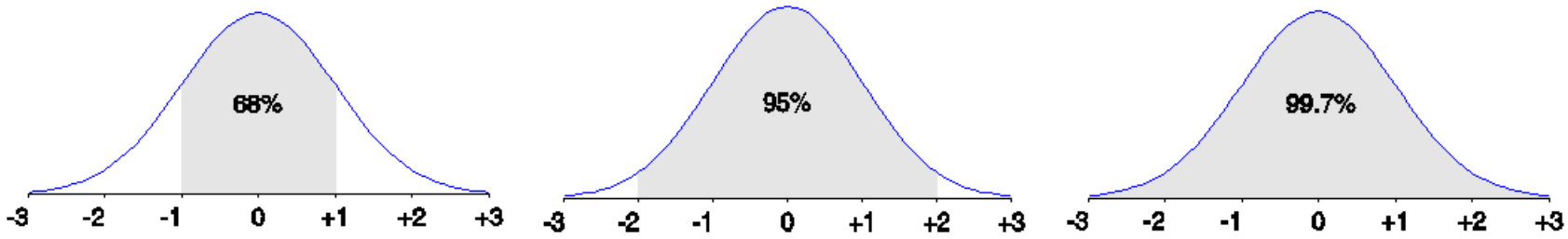
Measuring Dispersion of Data: Quartiles

- Suppose that set of observations for numeric attribute X is sorted in increasing order.
- **Quantiles** are points taken at regular intervals of a data distribution, dividing it into essentially equal size consecutive sets.
  - The **k^{th} q-quantile** for a given data distribution is the value x such that at most **k/q** of the data values are less than x and at most **(q-k)/q** of the data values are more than x, where k is an integer such that $0<k<q$. There are q-1 q-quantiles.
  - The 100-quantiles are more commonly referred to as **percentiles**; they divide the data distribution into 100 equal-sized consecutive sets.
- **Quartiles**: The 4-quantiles are the three data points that split the data distribution into four equal parts; each part represents one-fourth of the data distribution.



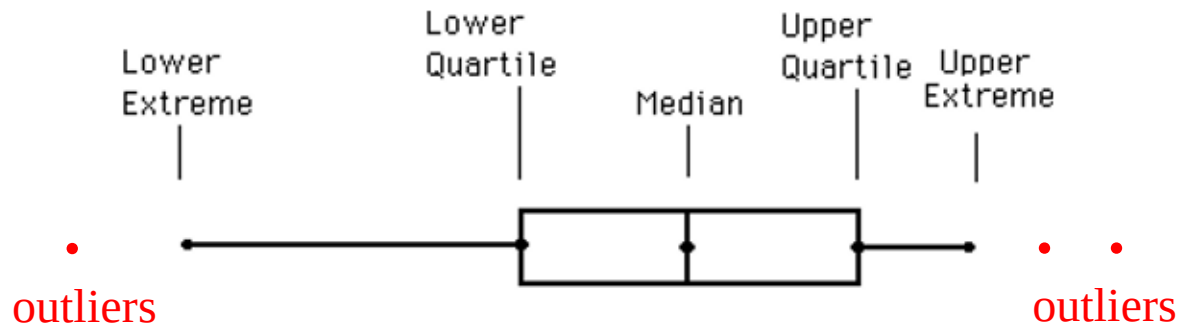
Measuring Dispersion of Data: Outliers

- **Outliers** can be identified by the help of ***interquartile range*** or ***standard deviation*** measures.
  - Suspected outliers are values falling at least 1.5xIQR above the third quartile or below the first quartile.
  - Suspected outliers are values that fall outside of the range of $\mu - N\sigma$ and $\mu + N\sigma$ where μ is mean and σ is standard deviation. N can be chosen as 2.5.
- The normal distribution curve: (μ: mean, σ: standard deviation)
  - From $\mu - \sigma$ to $\mu + \sigma$: contains about 68% of the measurements
  - From $\mu - 2\sigma$ to $\mu + 2\sigma$: contains about 95% of it
  - From $\mu - 3\sigma$ to $\mu + 3\sigma$: contains about 99.7% of it



Measuring Dispersion of Data: Boxplot Analysis

- **Five-number summary** of a distribution: Minimum, Q1, Median, Q3, Maximum
- **Boxplots** are a popular way of visualizing a distribution and a boxplot incorporates ***five-number summary***:
  - The ends of the box are at the **quartiles Q1 and Q3**, so that the box length is the interquartile range, IQR.
  - The **median** is marked by a line within the box. (**median** of values in IQR)
  - Two lines outside the box extend to the **smallest and largest observations** (outliers are excluded). Outliers are marked separately.
    - If there are no outliers, lower extreme line is the smallest observation (Minimum) and upper extreme line is the largest observation (Maximum).



Measuring Dispersion of Data: Example

Consider following two attribute values:

attr1: {2,3,4,5,6,7,8,9} **attr2: {1,5,9,10,11,12,18,30}**

Which attribute has biggest standard deviation? Do not compute standard deviations.

Give interquartile ranges of attribute values?

Are there any outliers (wrt IQR) in these datasets?

Give a 4 element dataset whose standard deviation is zero?

Measuring Dispersion of Data: Example

Consider following two attribute values:

attr1: {2,3,4,5,6,7,8,9} **attr2: {1,5,9,10,11,12,18,30}**

Which attribute has biggest standard deviation? Do not compute standard deviations.

attr2

Give interquartile ranges of attribute values?

attr1: Q1: (3+4)/2=3.5 Q3:(7+8)/2=7.5 IQR:3.5-7.5 = 4

attr2: Q1: (5+9)/2=7 Q3:(12+18)/2=15 IQR:7-15 = 8

Are there any outliers (wrt IQR) in these datasets?

Yes. 30 in attr2. 30 > 15+1.5*IQR

Give a 4 element dataset whose standard deviation is zero? **{1,1,1,1}**

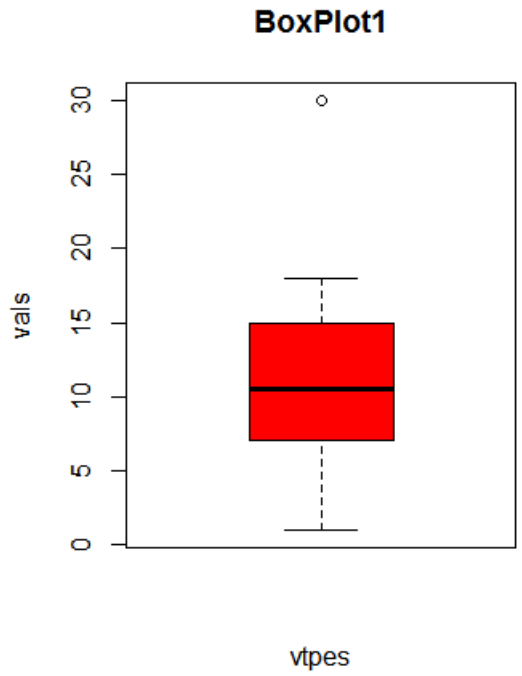
Graphic Displays of Basic Statistical Descriptions

- **Boxplot**: graphic display of five-number summary
- **Bar Chart:** compare data across different categories.
- **Histogram**: x-axis are values, y-axis represent frequencies
- **Quantile plot**: each value x_i is paired with f_i indicating that approximately 100 f_i % of data are $\leq x_i$
- **Quantile-quantile (q-q) plot**: graphs the quantiles of one univariant distribution against the corresponding quantiles of another.
- **Scatter plot**: each pair of values is a pair of coordinates and plotted as points in the plane.

Boxplot: Example in R

Boxplot: graphic display of five-number summary

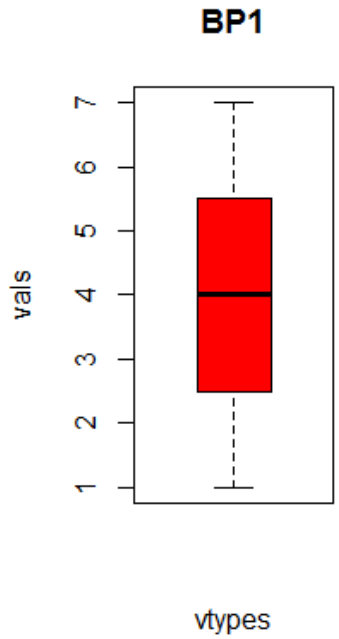
```
t3 <- data.frame(vals=c(1,5,9,10,11,12,18,30),vtypes=c(1,1,1,1,1,1,1,1))
boxplot(vals ~ vtypes, data = t3, col = "red", xlab="vtpes",
        ylab="vals",main="BoxPlot1")
```




```
fivenum(c(1,5,9,10,11,12,18,30))
1.0  7.0 10.5 15.0 30.0
```

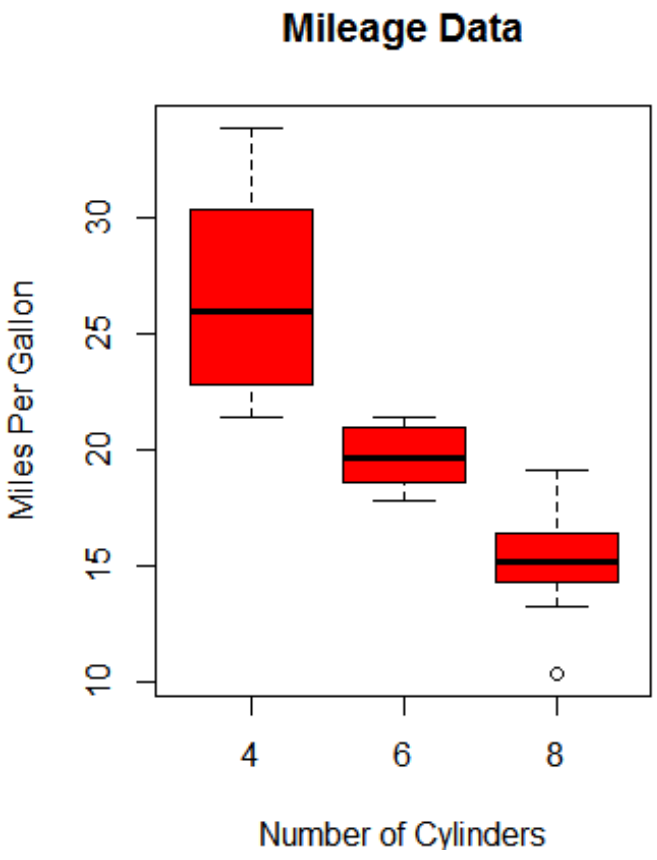
Boxplot: Example in R

```
t7 <- data.frame(vals=c(1,2,3,4,5,6,7),vtypes=c(1,1,1,1,1,1,1))
boxplot(vals~vtypes,data=t7,col="red",xlab="vtypes",ylab="vals",main="BP1")
fivenum(c(1,2,3,4,5,6,7))
1.0 2.5 4.0 5.5 7.0
```



Boxplot: Example in R

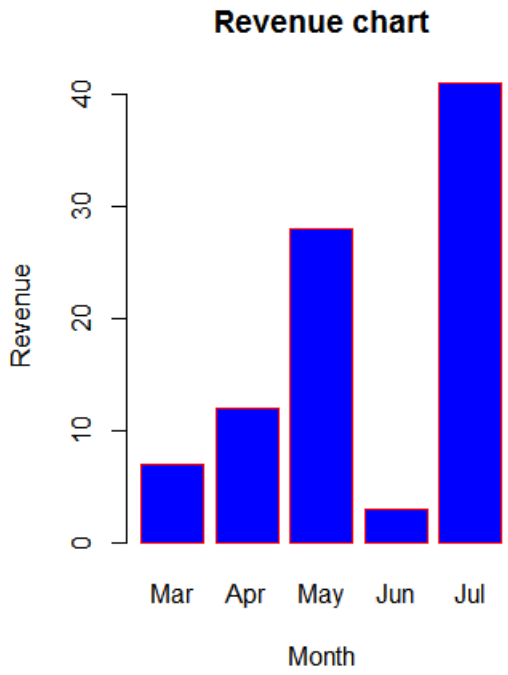
```
input <- mtcars[,c('mpg','cyl')]
boxplot(mpg ~ cyl, data = mtcars, xlab = "Number of Cylinders",
        ylab = "Miles Per Gallon", col="red", main = "Mileage Data")
```




```
> input
                  mpg cyl
Mazda RX4        21.0   6
Mazda RX4 Wag    21.0   6
Datsun 710       22.8   4
Hornet 4 Drive   21.4   6
Hornet Sportabout 18.7  8
Valiant          18.1   6
Duster 360       14.3   8
Merc 240D        24.4   4
Merc 230         22.8   4
Merc 280         19.2   6
Merc 280C        17.8   6
Merc 450SE       16.4   8
Merc 450SL       17.3   8
Merc 450SLC      15.2   8
Cadillac Fleetwood 10.4 8
Lincoln Continental 10.4 8
Chrysler Imperial 14.7  8
Fiat 128         32.4   4
Honda Civic      30.4   4
Toyota Corolla   33.9   4
Toyota Corona    21.5   4
Dodge Challenger 15.5   8
AMC Javelin      15.2   8
Camaro Z28       13.3   8
Pontiac Firebird 19.2   8
Fiat X1-9        27.3   4
Porsche 914-2    26.0   4
Lotus Europa     30.4   4
Ford Pantera L   15.8   8
Ferrari Dino     19.7   6
Maserati Bora    15.0   8
Volvo 142E       21.4   4
```

Bar Chart: Example in R

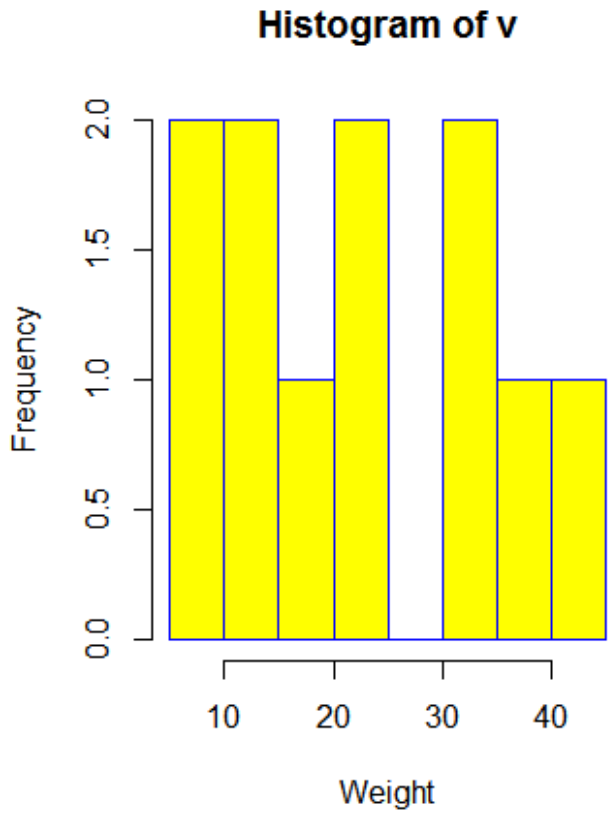
```
H <- c(7,12,28,3,41)
M <- c("Mar","Apr","May","Jun","Jul")
barplot(H,names.arg = M,xlab = "Month",ylab = "Revenue",col = "blue",
        main = "Revenue chart",border = "red")
```



- A **bar chart** represents data in rectangular bars with length of the bar proportional to the value of the variable.

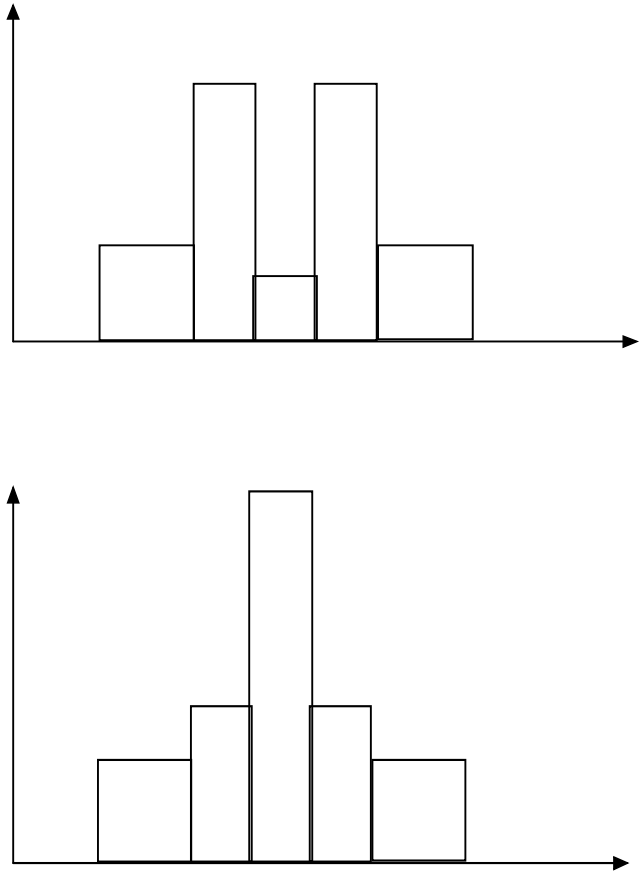
Histogram: Example in R

```
v <-  c(9,13,21,8,36,22,12,41,31,33,19)
hist(v,xlab = "Weight",col = "yellow",border = "blue")
```



- A **histogram** represents the frequencies of values of a variable bucketed into ranges.
- Histogram is similar to bar chat but the difference is it groups the values into continuous ranges.

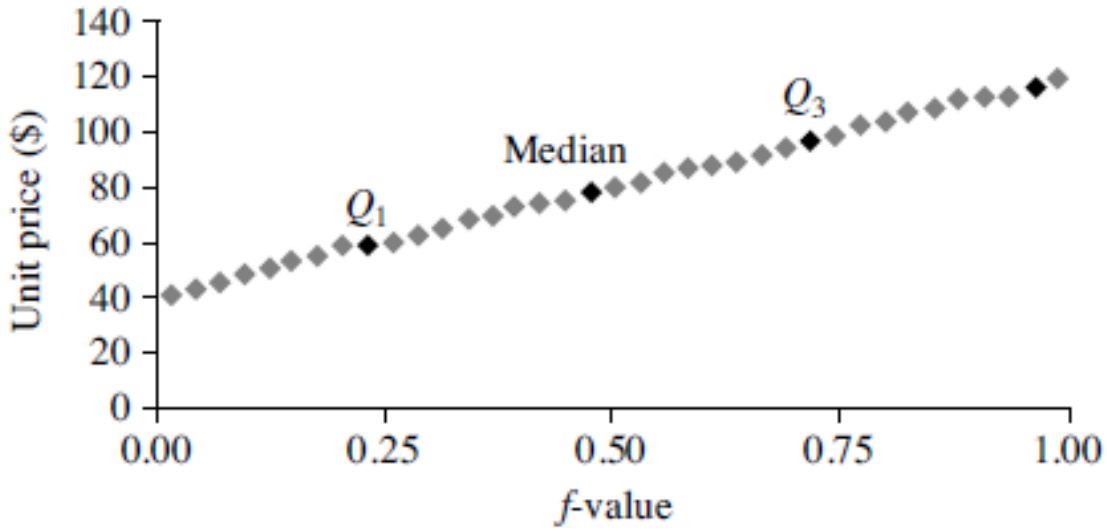
Histograms Often Tell More than Boxplots



- The two histograms shown in the left may have the same boxplot representation
  - The same values for: min, Q1, median, Q3, max
- But they have rather different data distributions

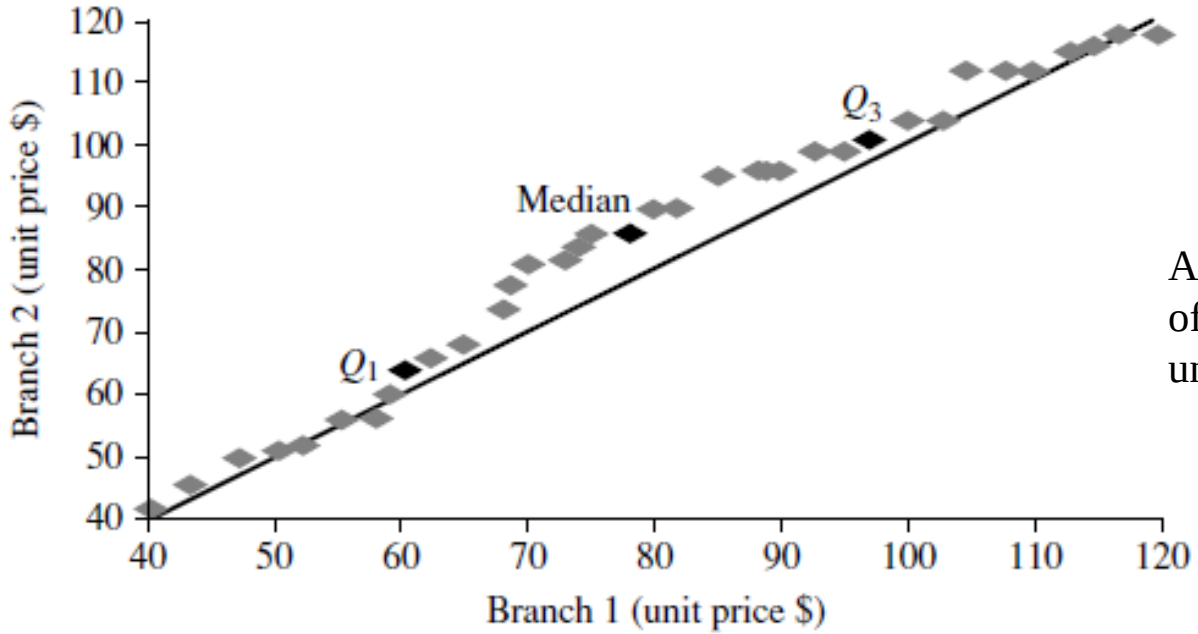
Quantile Plot

- Displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences)
- Plots **quantile** information
  - For a data x_i data sorted in increasing order, f_i indicates that approximately 100 f_i% of the data are below or equal to the value x_i



Quantile-Quantile (Q-Q) Plot

- Graphs the quantiles of one univariate distribution against the corresponding quantiles of another
- View: Is there is a shift in going from one distribution to another?
- Example shows unit price of items sold at Branch 1 vs. Branch 2 for each quantile. Unit prices of items sold at Branch 1 tend to be lower than those at Branch 2.



A straight line that represents the case of when, for each given quantile, the unit price at each branch is the same.

Scatter Plot

- A **scatter plot** is one of the most effective graphical methods for determining if there appears to be a relationship, pattern, or trend between two numeric attributes.
  - To construct a scatter plot, each pair of values is treated as a pair of coordinates in an algebraic sense and plotted as points in the plane.

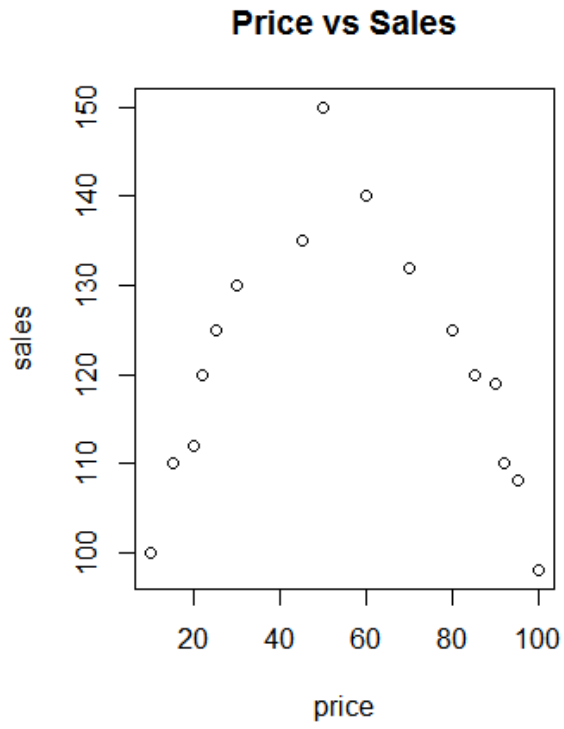
- The **scatter plot** is a useful method for providing a first look at bivariate data to see clusters of points and outliers, or to explore the possibility of correlation relationships.

- Two attributes, X, and Y, are **correlated** if one attribute implies the other.
- **Correlations** can be **positive**, **negative**, or null (**uncorrelated**).

Scatter Plot: Example in R

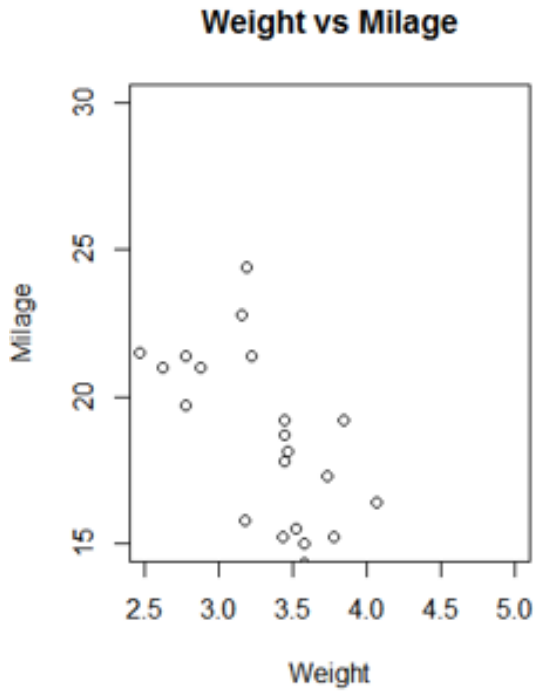
```
t2 <- data.frame(price=c(10,15,20,22,25,30,45,50,60,70,80,85,90,92,95,100),
   sales=c(100,110,112,120,125,130,135,150,140,132,125,120,119,110,108,98))

plot(x=t2$price,y=t2$sales,xlab="price",ylab="sales",main="Price vs Sales")
```



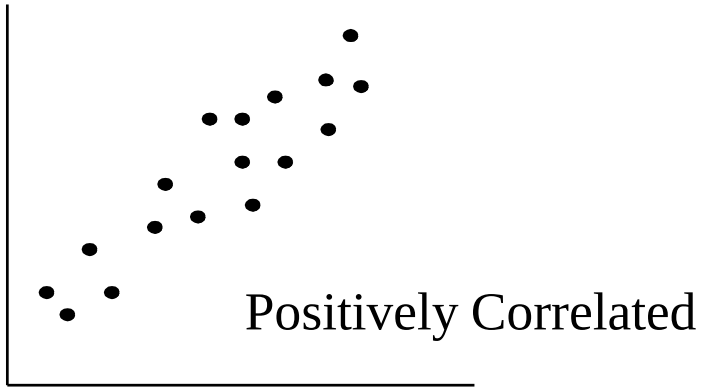
Scatter Plot: Example in R

```
input <- mtcars[,c('wt','mpg')]
plot(x = input$wt,y = input$mpg,
     xlab = "Weight",
     ylab = "Milage",
     xlim = c(2.5,5),
     ylim = c(15,30),
     main = "Weight vs Milage"
)
```

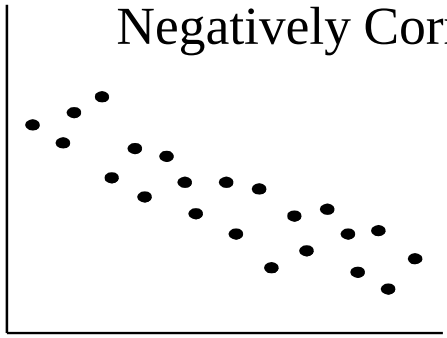


- Scatterplots show many points plotted in the Cartesian plane. Each point represents the values of two variables.
- One variable is chosen in the horizontal axis and another in the vertical axis.

Scatter Plot:
Positively and Negatively Correlated Data

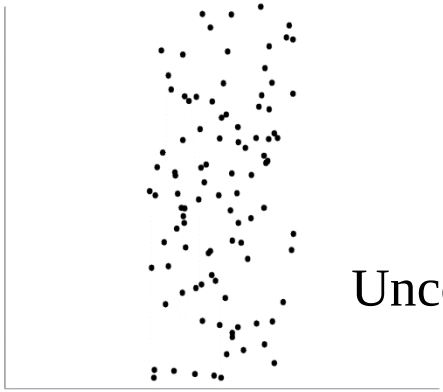


Positively Correlated



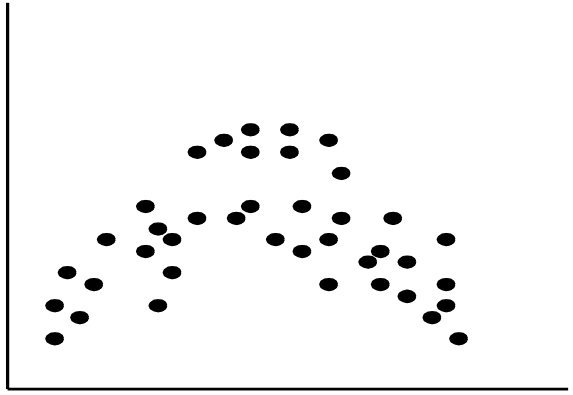
Negatively Correlated

Uncorrelated



Uncorrelated

Scatter Plot:
Positively and Negatively Correlated Data



The left half fragment is positively correlated

The right half is negative correlated

- Data Objects and Attribute Types
- Basic Statistical Descriptions of Data
- **Measuring Data Similarity and Dissimilarity**

Similarity and Dissimilarity

Similarity

- The similarity between two objects is a numerical measure of the degree to which the two objects are alike.
- Similarities are higher for pairs of objects that are more alike.
- Similarities are usually non-negative and are often between 0 (no similarity) and 1 (complete similarity).

Dissimilarity

- The dissimilarity between two objects is a numerical measure of the degree to which the two objects are different.
- Dissimilarities are lower for more similar pairs of objects.
- The term **distance** is used as a synonym for dissimilarity, although the distance is often used to refer to a special class of dissimilarities.
- Dissimilarities sometimes fall in the interval [0,1], but it is also common for them to range from 0 to ∞.

Proximity refers to a similarity or dissimilarity

Similarity/Dissimilarity for Simple Attributes

- The proximity of objects with a number of attributes is typically defined by combining the proximities of individual attributes.
- Consider objects described by one nominal attribute.
  - What would it mean for two such objects to be similar?

- *p* and *q* are the attribute values for two data objects

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
Ordinal	$d = \frac{\lvert p-q \rvert}{n-1}$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - \frac{\lvert p-q \rvert}{n-1}$
Interval or Ratio	$d = \lvert p - q \rvert$	$s = -d$, $s = \frac{1}{1+d}$ or $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Dissimilarities between Data Objects
Distance on Numeric Data: Euclidean Distance

- Assume that our data objects have n numeric attributes.

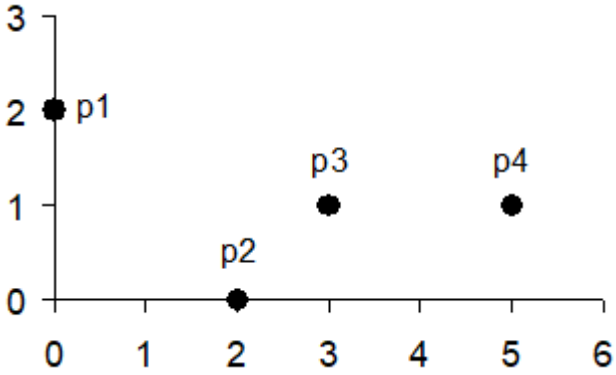
Euclidean Distance

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

where n is the number of attributes and x_k and x_k are k^{th} attributes of data objects x and y.

- Normally attributes are numeric.
- Standardization is necessary, if scales of attributes differ.

Distance on Numeric Data: Euclidean Distance



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Distance Matrix

Distance on Numeric Data: Minkowski Distance

- **Minkowski Distance** is a generalization of Euclidean Distance

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{1/r}$$

where

  - r is a parameter,
  - n is the number of attributes and
  - x_k and y_k are kth attributes of data objects x and x.

- Note that Minkowski Distance is Euclidean Distance when r=2

Distance on Numeric Data: Minkowski Distance

- **r = 1** : **Manhattan (**city block , taxicab, **L_1 norm**) **distance**.
  - A common example of this is the Hamming distance, which is just the number of bits that are different between two binary vectors

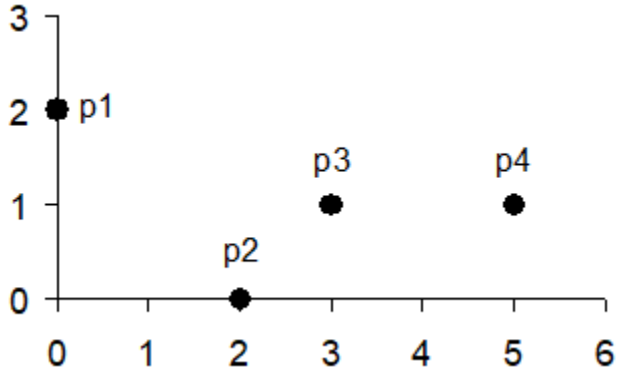
- **r = 2** : **Euclidean distance** (**L_2 norm**)

- **$r \rightarrow \infty$** : **Supremum** (**L_{max}** or **L_∞ norm**) **distance**.
  - This is the maximum difference between any component of the vectors

$$d(\mathbf{x}, \mathbf{y}) = \lim_{r \to \infty} \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{1/r}$$

- Do not confuse *r* with *n*, i.e., all these distances are defined for all numbers of attributes.

Distance on Numeric Data: Minkowski Distance



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

L_1	p1	p2	p3	p4
p1	0.0	4.0	4.0	6.0
p2	4.0	0.0	2.0	4.0
p3	4.0	2.0	0.0	2.0
p4	6.0	4.0	2.0	0.0

Manhattan (L_1) distance matrix

	p1	p2	p3	p4
p1	0.0	2.8	3.2	5.1
p2	2.8	0.0	1.4	3.2
p3	3.2	1.4	0.0	2.0
p4	5.1	3.2	2.0	0.0

Euclidean (L_2) distance matrix

L_∞	p1	p2	p3	p4
p1	0.0	2.0	3.0	5.0
p2	2.0	0.0	1.0	3.0
p3	3.0	1.0	0.0	2.0
p4	5.0	3.0	2.0	0.0

Supremum (L_∞) distance matrix

Properties of Distances

- Distances, such as the Euclidean distance, have some well-known properties.
- If distance d(x, y) between x and y, then the following properties hold.

1. **Positivity**
   a) **d(x, y) ≥ 0** for all **x** and **y**,
   b) **d(x, y) = 0** only if **x** = **y**.
2. **Symmetry**
   d(x, y) = d(y, x) for all **x** and **y**.
3. **Triangle Inequality**
   d(x, z) ≤ d(x, y) + d(y, z) for all points **x**, **y**, and **z**.

- Measures that satisfy all three properties are known as **metrics**.
- Some dissimilarities do not satisfy one or more of the metric properties.
  - Examples: set difference, time difference

Properties of Similarities

- For similarities, *triangle inequality* typically does not hold, but *symmetry* and *positivity* typically do.

- If s(x, y) is the similarity between points x and y, then the typical properties of similarities are:

1. **Positivity**

   ***s*(x, y) = 1** only if **x** = **y**. $(0 \leq s(x,y) \leq 1)$

2. **Symmetry**

   ***s*(x, y) = *s*(y, x)** for all **x** and **y**.

Proximity Measures for Binary Attributes

- Let **x** and **y** be two data objects that consist of *n* binary attributes.
- The comparison of two such objects, i.e., two binary vectors, leads to the following four frequencies:

  f_{00} = the number of attributes where **x** is 0 and **y** is 0

  f_{01} = the number of attributes where **x** is 0 and **y** is 1

  f_{10} = the number of attributes where **x** is 1 and **y** is 0

  f_{11} = the number of attributes where **x** is 1 and **y** is 1

- **Distance measure for symmetric binary variables**:

$$\mathbf{d(x, y) = \frac{f_{01} + f_{10}}{f_{00} + f_{01} + f_{10} + f_{11}}}$$

- **Distance measure for asymmetric binary variables**:

$$\mathbf{d(x, y) = \frac{f_{01} + f_{10}}{f_{01} + f_{10} + f_{11}}}$$

Similarity Between Binary Vectors: Simple Matching and Jaccard Coefficients

- Similarity measures between objects that contain only binary attributes are called **similarity coefficients**, and typically have values between 0 and 1.

- **Simple Matching Coefficient (SMC)** counts both presences and absences equally and it is normally used for *symmetric binary attributes*.

$$\mathbf{SMC} = \frac{\mathbf{number\ of\ matching\ attributes}}{\mathbf{number\ of\ attributes}} = \frac{\mathbf{f_{11} + f_{00}}}{\mathbf{f_{00} + f_{01} + f_{10} + f_{11}}}$$

- **Jaccard Coefficient (J)** counts only presences and it is frequently for *asymmetric binary attributes.*

$$\mathbf{J} = \frac{\mathbf{number\ of\ 11\ matches}}{\mathbf{number\ of\ not-both-zero\ attributes}} = \frac{\mathbf{f_{11}}}{\mathbf{f_{01} + f_{10} + f_{11}}}$$

SMC versus Jaccard Coefficient: Example

p = 1 0 0 0 0 0 0 0 0 0
q = 0 0 0 0 0 0 1 0 0 1

f_{01} = 2 (the number of attributes where p was 0 and q was 1)
f_{10} = 1 (the number of attributes where p was 1 and q was 0)
f_{00} = 7 (the number of attributes where p was 0 and q was 0)
f_{11} = 0 (the number of attributes where p was 1 and q was 1)

SMC = $(f_{11} + f_{00}) / (f_{01} + f_{10} + f_{11} + f_{00})$ = (0+7) / (2+1+0+7) = 0.7

J = $(f_{11}) / (f_{01} + f_{10} + f_{11})$ = 0 / (2 + 1 + 0) = 0

Cosine Similarity

- A **document** can be represented by thousands of attributes, each recording the frequency of a particular word or phrase in the document.

Document	*team*	*coach*	*hockey*	*baseball*	*soccer*	*penalty*	*score*	*win*	*loss*	*season*
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

- A similarity measure for documents needs to ignores 0–0 matches like the Jaccard measure, but also must be able to handle non-binary vectors.
- **Cosine similarity** is one of the most common measure of document similarity.

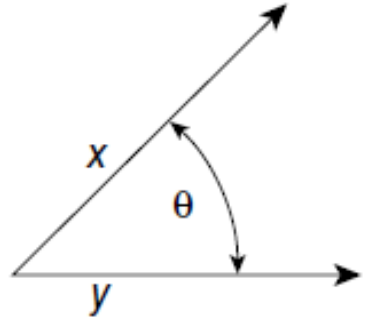
$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \, \|\mathbf{y}\|}$$

where x and y are two document vectors

- where • indicates vector dot product $\mathbf{x} \cdot \mathbf{y} = \sum_{k=1}^{n} x_k y_k$ and $\|\mathbf{x}\|$ is the length of vector **x**. $\|\mathbf{x}\| = \sqrt{\sum_{k=1}^{n} x_k^2} = \sqrt{\mathbf{x} \cdot \mathbf{x}}$

Cosine Similarity

- **Cosine similarity** really is a measure of the (cosine of the) angle between x and y.
  - If the cosine similarity is 1, the angle between x and y is 0^o, and x and y are same
  - If the cosine similarity is 0, then the angle between x and y is 90^o, and they do not share any terms.



- Cosine similarity can be written as

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}}{\|\mathbf{x}\|} \cdot \frac{\mathbf{y}}{\|\mathbf{y}\|}$$

  - Dividing x and y by their lengths normalizes them to have a length of 1.
  - This means that cosine similarity does not take the magnitude of the two data objects into account when computing similarity.
- Euclidean distance might be a better choice when magnitude is important.

Cosine Similarity : Example

x = (3, 2, 0, 5, 0, 0, 0, 2, 0, 0)

y = (1, 0, 0, 0, 0, 0, 0, 1, 0, 2)

$$x \cdot y = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$$
$$\|x\| = \sqrt{3*3 + 2*2 + 0*0 + 5*5 + 0*0 + 0*0 + 0*0 + 2*2 + 0*0 + 0*0} = 6.48$$
$$\|y\| = \sqrt{1*1 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 1*1 + 0*0 + 2*2} = 2.24$$

cos(x,y) = (x.y) / (||x|| ||y||) = 5 / (6.48*2.24) = 0.34

Extended Jaccard Coefficient

- **Extended Jaccard Coefficient** can be used for document data and that reduces to the Jaccard coefficient in the case of binary attributes.
- Extended Jaccard Coefficient is also known as Tanimoto coefficient.

$$EJ(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - \mathbf{x} \cdot \mathbf{y}}$$

Correlation

- **Correlation** measures the linear relationship between objects
- **Pearson's correlation coefficient** between two data objects, x and y:

$$\text{corr}(\mathbf{x}, \mathbf{y}) = \frac{\text{covariance}(\mathbf{x}, \mathbf{y})}{\text{standard\_deviation}(\mathbf{x}) * \text{standard\_deviation}(\mathbf{y})} = \frac{s_{xy}}{s_x \; s_y}$$

- where

$$\text{covariance}(\mathbf{x}, \mathbf{y}) = s_{xy} = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})(y_k - \overline{y})$$

$$\text{standard\_deviation}(\mathbf{x}) = s_x = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})^2}$$

$$\text{standard\_deviation}(\mathbf{y}) = s_y = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (y_k - \overline{y})^2}$$

$$\overline{x} = \frac{1}{n} \sum_{k=1}^{n} x_k \text{ is the mean of x}$$

$$\overline{y} = \frac{1}{n} \sum_{k=1}^{n} y_k \text{ is the mean of y}$$

Correlation: Perfect Correlation

- Correlation is always in the range -1 to 1.
- A correlation of 1 (-1) means that x and y have a perfect positive (negative) linear relationship.

- A perfect negative linear relationship (correlation: -1)

  x = (-3, 6, 0, 3, -6) s_{xy} = -7.5 s_x = 4.74341649 s_y = 1.58113883

  y = (1, -2, 0, -1, 2) corr(x,y) = -1

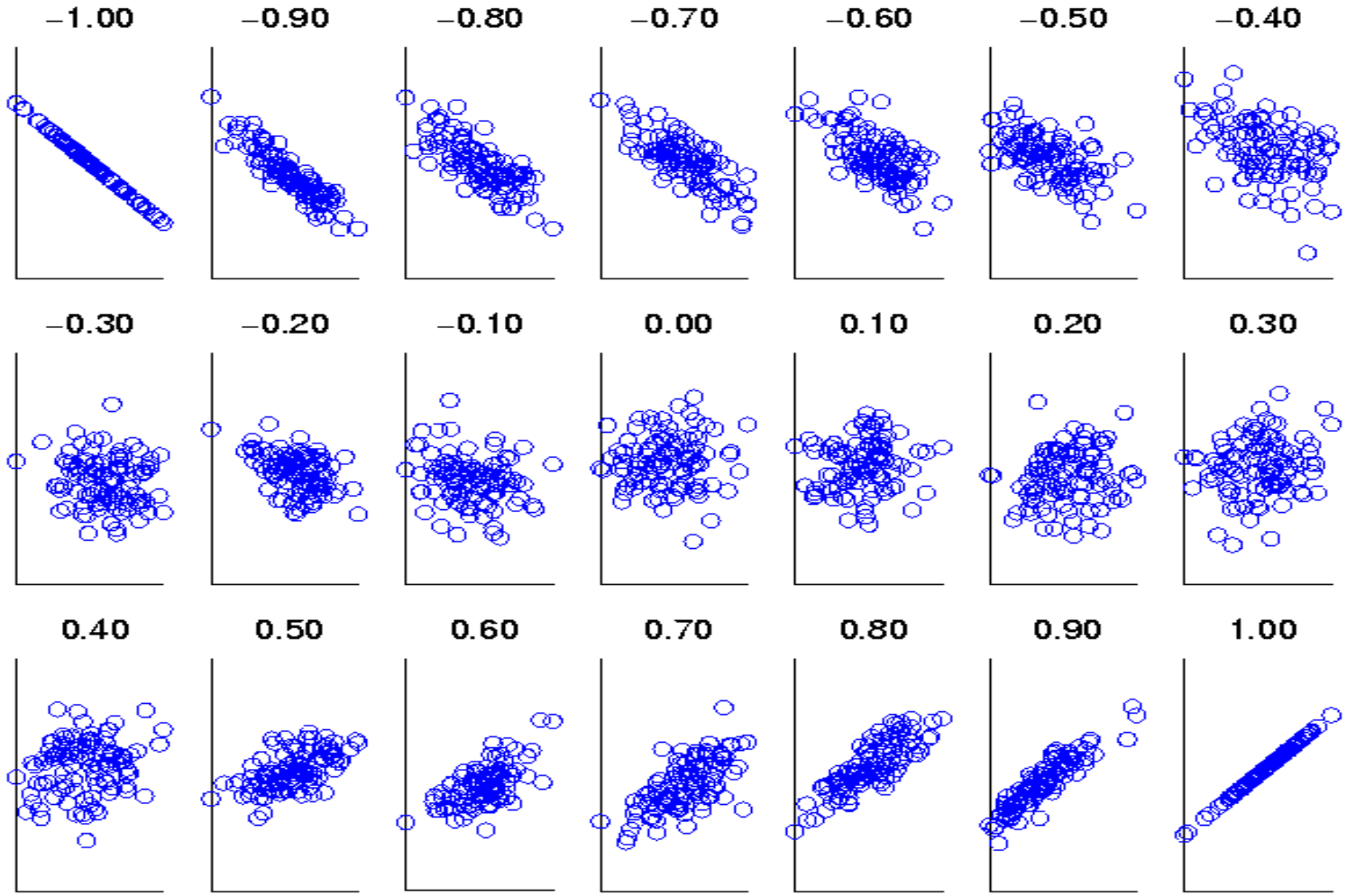
- A perfect positive linear relationship (correlation: +1)

  x = (3 , 6 , 0 , 3 , 6) s_{xy} = 2.1 s_x = 2.50998008 s_y = 0.836660027

  y = (1 , 2 , 0 , 1 , 2) corr(x,y) = +1

Visually Evaluating Correlation

scatter plots showing the similarity from –1 to 1



Drawback of Correlation: Non-linear Relationships

- If the correlation is 0, then there is no linear relationship between the attributes of the two data objects.
- However, non-linear relationships may still exist. In the following example, $y_i = x_i^2$, but their correlation is 0.

  x = (−3, −2, −1, 0, 1, 2, 3)

  y = (9, 4, 1, 0, 1, 4, 9)

- mean(x) = 0 mean(y) = 4 std(x) = 2.16 std(y) = 3.74

$$\text{corr}(\mathbf{x}, \mathbf{y}) = \frac{\text{covariance}(\mathbf{x}, \mathbf{y})}{\text{standard\_deviation}(\mathbf{x}) * \text{standard\_deviation}(\mathbf{y})} = \frac{s_{xy}}{s_x \; s_y}$$

$$\text{covariance}(\mathbf{x}, \mathbf{y}) = s_{xy} = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})(y_k - \overline{y})$$

corr(x,y) = (-3)(5)+(-2)(0)+(-1)(-3)+(0)(-4)+(1)(-3)+(2)(0)+3(5) / (6 * 2.16 * 3.74)

= 0

Issues in Proximity Calculation

Important issues related to proximity measures:

1. How to handle the case in which attributes have different scales and/or are correlated
   - This situation is often described by saying that "the variables have different scales."
   - Example: Euclidean distance is used to measure the distance between people based on two attributes: age and income.
     - Unless these attributes are standardized, distance between two people is dominated by income.
     - We have to both attributes have same range (Ex: 0 – 1) ➔ **Normalization**
2. How to calculate proximity between objects that are composed of different types of attributes, e.g., quantitative and qualitative,
3. How to handle proximity calculation when attributes have different weights; i.e., when not all attributes contribute equally to the proximity of objects.

General Approach for Combining Similarities

- Sometimes attributes are of many different types, but an overall similarity is needed.
- The following algorithm is effective for computing an overall similarity between two heterogeneous objects, x and y, with different types of attributes.

1: For the k^{th} attribute, compute a similarity, $s_k(x,y)$, in the range [0,1].

2: Define an indicator variable, δ_k, for the k^{th} attribute as follows:

$$\delta_k = \begin{cases} 0 & \text{if the } k^{th} \text{ attribute is an asymmetric attribute and} \\ & \text{both objects have a value of 0, or if one of the objects} \\ & \text{has a missing value for the } k^{th} \text{ attribute} \\ 1 & \text{otherwise} \end{cases}$$

3: Compute the overall similarity between the two objects using the following formula:

$$\text{similarity}(x, y) = \frac{\sum_{k=1}^{n} \delta_k s_k(x, y)}{\sum_{k=1}^{n} \delta_k}$$

Using Weights to Combine Similarities

- We may not want to treat all attributes the same.
- Use weights w_k which are between 0 and 1 and they sum to 1.

$$\text{similarity}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k=1}^{n} w_k \delta_k s_k(\mathbf{x}, \mathbf{y})}{\sum_{k=1}^{n} \delta_k}$$

Modified Minkowski distance

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^{n} w_k |x_k - y_k|^r \right)^{1/r}$$

Selecting the Right Proximity Measure

- For many types of dense, continuous data, metric distance measures such as Euclidean distance are often used.
- The cosine, Jaccard, and extended Jaccard measures are appropriate for sparse, asymmetric data, most objects have only a few of the characteristics described by the attributes and thus, are highly similar in terms of the characteristics they do not have.
- In some cases, transformation or normalization of the data is important for obtaining a proper similarity measure since such transformations are not always present in proximity measures.

- **The proper choice of a proximity measure can be a time-consuming task that requires careful consideration of both domain knowledge and the purpose for which the measure is being used.**
- **A number of different similarity measures may need to be evaluated to see which ones produce results that make the most sense**.