

IOT607U Data Mining

Week 2: Data

Dr Lin Wang

School of EECS, Queen Mary University of London

Last week: Introduction

- Preliminaries
- Why data mining?
- What is data mining?
- Data mining tasks
- Challenges in data mining

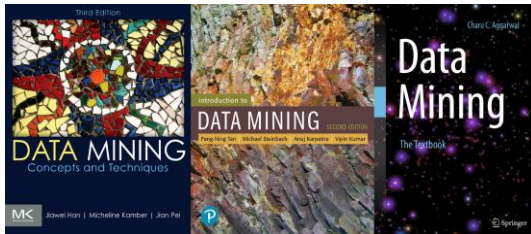
This week's contents

1. Attributes and Objects
2. Characteristics of Data
3. Data Representations
4. Basic Statistical Descriptions of Data
5. Similarity and Distance



Reading

- Chapter 2 of J. Han, M. Kamber, J. Pei, “Data Mining: Concepts and Techniques”, 3rd edition, Elsevier/Morgan Kaufmann, 2012
- Chapter 2 of P.-N. Tan, M. Steinbach, A. Karpatne, V. Kumar, “Introduction to Data Mining”, 2nd edition, Pearson, 2019
- Chapter 3 of C. C. Aggarwal, “Data Mining: The Textbook”, Springer, 2015



Attributes and Objects

Getting to Know Your Data

A **dataset** can be viewed as a collection of **data objects**. Other names for data objects are *data points*, *samples*, *observations*, *vectors*, *records*, *patterns*, *events*, *cases*, *examples*, *instances*, and *entities*.

Data objects are described by a number of **attributes** that capture their basic characteristics. Other names for attributes are *features*, *variables*, *characteristics*, *fields*, and *dimensions*.

Attributes

Tid	Attributes			
	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

Objects

What Is an Object?

A **data object** represents an **entity**. If the data objects are stored in a database, they are *data tuples*: rows of a database correspond to data objects, and columns correspond to attributes.

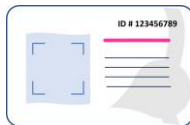
- e.g. in a medical database, the objects can be patients
- e.g. in a university database, the objects can be students, professors, and courses
- e.g. in a sales database, the objects may be customers, store items, and sales

What Is an Attribute?

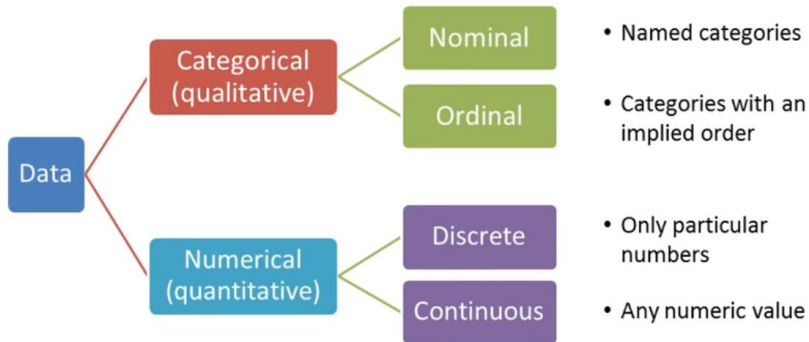
An **attribute** is a **property**, **characteristic**, or **feature of an object** that may vary, either from one object to another or from one time to another.

- e.g. attributes describing a customer object can include *customer ID*, *name*, and *address*

A set of attributes used to describe a given object is called an *attribute vector*.



TYPES OF ATTRIBUTES (DATA VARIABLES)



TYPES OF VARIABLES

Chart of Nominal data

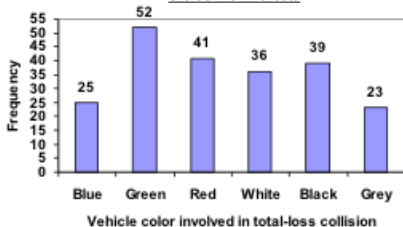
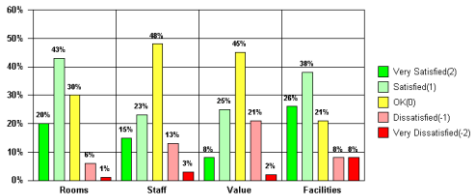


Chart of Ordinal data



Nominal Attributes

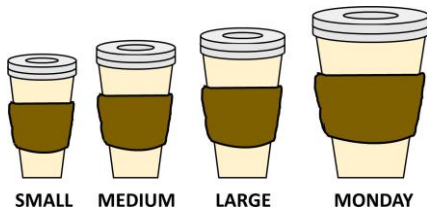
Nominal means “relating to names.” The values of a nominal attribute are symbols or names of things, which do not have a meaningful order. Each value represents some kind of category, code, or state.

Although nominal attributes are not quantitative, it is possible to represent such symbols or “names” with numbers.

However, mathematical operations on values of nominal attributes are not meaningful. For example, it makes no sense to subtract one customer ID number from another.

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.



Discrete versus Continuous Attributes

A **discrete attribute** has a finite or countably infinite set of values, which may or may not be represented as integers. Such attributes can be categorical or numeric. Binary attributes are discrete.

A **continuous attribute** is one whose values are real numbers. Continuous attributes are typically represented as floating-point variables. Practically, real values can only be measured and represented with limited precision.

Characteristics of Data

Characteristics of Data

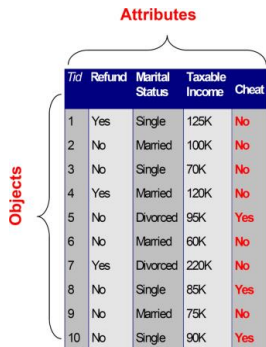
As we will discuss later, there are many types of datasets. However, few characteristics apply to many datasets and have a significant impact on the data mining techniques that are used:

- **Dimensionality**: high dimensional data brings a number of challenges
- **Sparsity**: only presence counts
- **Resolution**: patterns depend on the scale
- **Size**: type of analysis may depend on size of data

Dimensionality

The **dimensionality** of a data set is the number of attributes that the objects in the data set possess.

There exist difficulties associated with analyzing high-dimensional data; an important motivation in preprocessing the data is *dimensionality reduction*.



<i>Tid</i>	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

Sparsity

The **sparsity** of a data set refers to any data set with asymmetric features (i.e., not all features are equally important) which has a very large group of zero values and very little non-zero values.

In practical terms, sparsity is an advantage because usually only the non-zero values need to be stored and manipulated. This results in significant savings with respect to computation time and storage. Furthermore, some data mining algorithms work well only for sparse data.

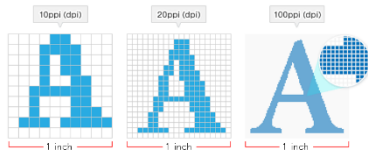
Dense											Sparse										
1	2	31	2	9	7	34	22	11	5		1	.	3	.	9	.	3
11	92	4	3	2	2	3	3	2	1		11	.	4	2	1	.
3	9	13	8	21	17	4	2	1	4		.	.	1	.	.	.	4	.	1	.	.
8	32	1	2	34	18	7	78	10	7		8	.	.	.	3	1
9	22	3	9	8	71	12	22	17	3		.	.	.	9	.	.	1	.	17	.	.
13	21	21	9	2	47	1	81	21	9		13	21	.	9	2	47	1	81	21	9	.
21	12	53	12	91	24	81	8	91	2	
61	8	33	82	19	87	16	3	1	55		19	8	16	.	.	55	.
54	4	78	24	18	11	4	2	99	5		54	4	.	.	.	11

Resolution

Data sets can have different level of **resolutions** and the data attributes can be different at different resolutions.

For instance, the surface of the Earth seems very uneven at a resolution of a few meters, but is relatively smooth at a resolution of tens of kilometers. The characteristics of the data depend on the level of resolution. If the resolution is too fine, a characteristic may not be visible or may be buried in noise; if the resolution is too coarse, the characteristic may disappear.

For example, weather predictions in the scale of hours vs months.



Data sets can have different **sizes**, i.e. they are captured at different scales.

Due to the increasing sizes of the data in modern-day applications, *their analysis in different scales* is an important concern in many data mining applications. There are two important scenarios for scalability:

- The data is stored on one or more machines, but they are too large to process efficiently.
- The data is generated continuously over time in high volume, and it is not practical to store it entirely. This scenario is that of *data streams*, in which the data need to be processed with the use of an online approach.

Data Representations

- **Basic Records**
- Graphs
- Ordered

Basic Records

In the basic records case there is no explicit relationship among data objects (records), and every data object has the same set of attributes.

<i>Tid</i>	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

Basic Record: Data Matrix

If the data objects in a collection of data all have the same fixed set of numeric attributes, then the data objects can be thought of as points (vectors) in a multidimensional space, where each dimension represents a distinct attribute describing the object:

- A data set represented by an $m \times n$ matrix, where there are m rows, one for each object, and n columns, one for each attribute

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

Basic Record: Transaction Data

Transaction Data is a special type of basic record data, where each transaction (record) involves a set of items.

- e.g. a set of products purchased by a customer during one shopping trip constitutes a transaction, while the individual products that were purchased are the items

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Basic Record: Document Data

If the order of the terms (words) in a document is ignored, then a document can be represented as a *term vector*, where each term is a component (attribute) of the vector and the value of each component is the frequency of the term (i.e., number of times that the corresponding term occurs in the document). This representation of a collection of documents is often called a **document-term matrix**.

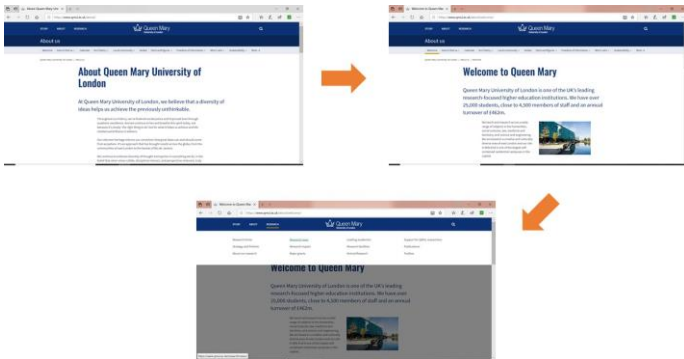
	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Data Representations

- Basic Records
- **Graphs**
- Ordered

Graph: World Wide Web

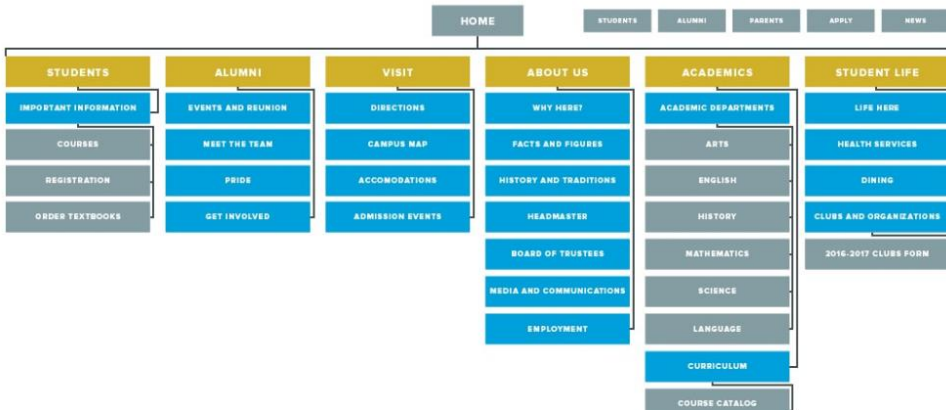
Data with Relationships (hyperlinks) among Objects



- e.g. QMUL webpage

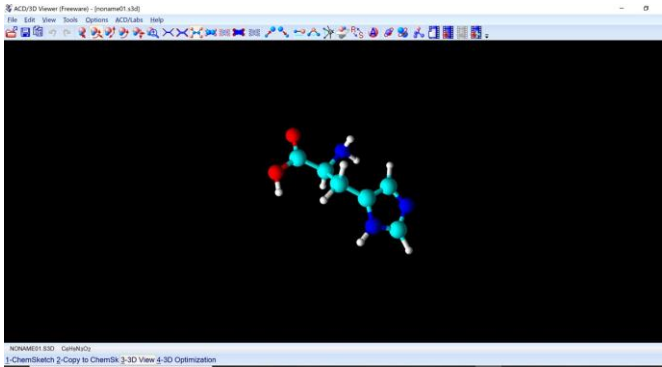
Graph: World Wide Web

Data with Relationships (hyperlinks) among Objects



Graph: Molecular Structures

Data with Objects That Are Graphs



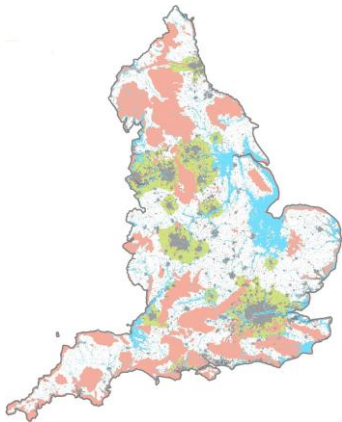
- e.g. molecular structure - L-histidine

Data Representations

- Basic Records
- Graphs
- **Ordered**

Ordered: Spatial Data

In the case of **spatial data**, some objects have spatial attributes, such as position or areas, as well as other types of attributes.

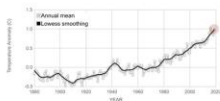


Ordered: Temporal Data

In **temporal (or sequential) data** attributes have relationships that involve order in time or space.

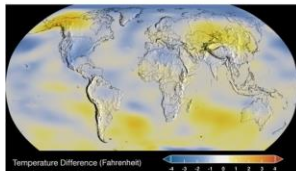
GLOBAL LAND-OCEAN TEMPERATURE INDEX

Data source: NASA's Goddard Institute for Space Studies (GISS)
Credit: NASA/GISS



TIME SERIES: 1884 TO 2019

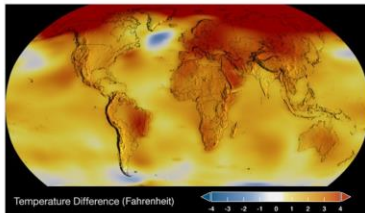
Data source: NASA/GISS
Credit: NASA Scientific Visualization Studio



► 1884 ◯ 2019

TIME SERIES: 1884 TO 2019

Data source: NASA/GISS
Credit: NASA Scientific Visualization Studio



► 1884 ◯ 2019

2019

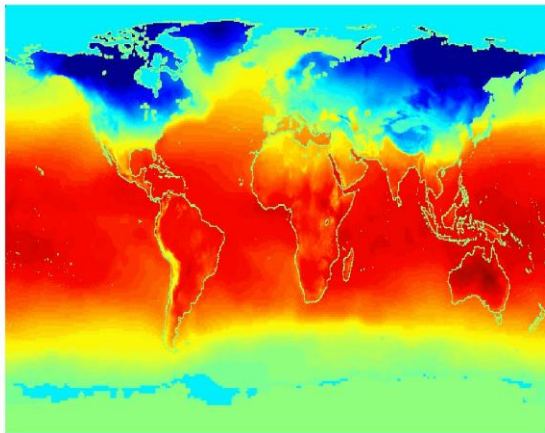
1979

- e.g. global land-ocean temperature changes in time

Ordered: Spatial-Temporal Data

**Average
Monthly
Temperature of
land and ocean**

Jan



Ordered: Sequence Data

Sequence data consists of a data set that is a sequence of individual entities, such as a sequence of words or letters.

**GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG**

- e.g. genetic information of plants and animals

Basic Statistical Descriptions of Data

- **Probability density function (PDF):**

- the probability that a discrete random variable is exactly **equal to some value**

$$P[X=x]$$

- **Cumulative distribution function (CDF):**

- the probability that X will take a value **less than or equal to x**

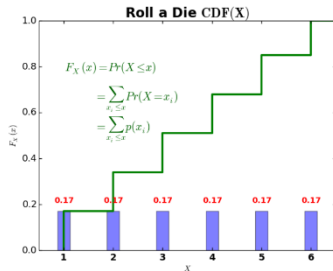
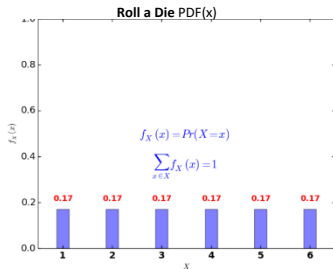
$$P[X \leq x]$$

Example: PDF

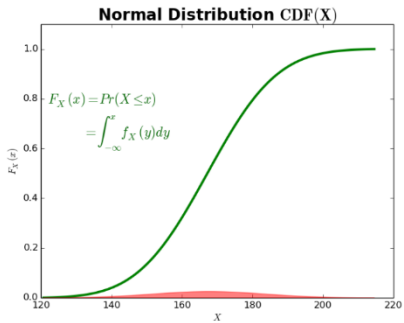
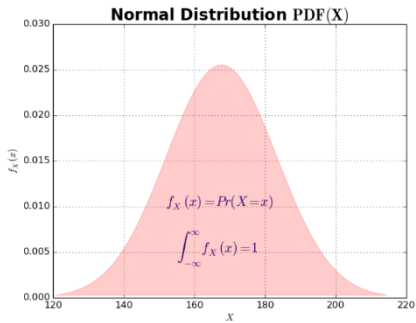
A dice has 6 values: 1-6. You get one value when you roll a dice.
Draw the PDF and CDF of the obtained value.

Example: PDF

A dice has 6 values: 1-6. You get one value when you roll a dice.
Draw the PDF and CDF of the obtained value.

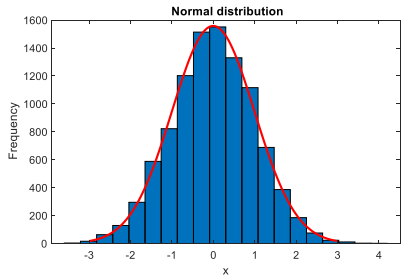


Example: PDF

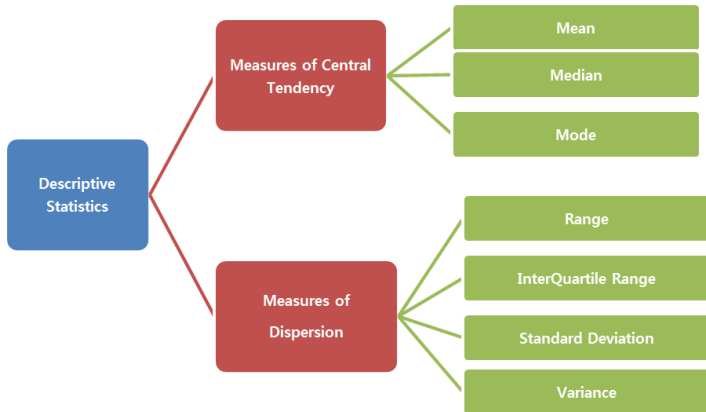


Terminology

- Histogram
 - Approximate representation of the distribution of numerical data
 - Bin (interval)
 - Frequency (number of occurrences)



Descriptive Statistics



CENTRALITY

mean, median, mode

Central Tendency

- With **centrality**, we aim to capture the **central tendency of a variable** or distribution of values.

Variable1 = [5, 5, 5, 5, 5] → centrality = ? 5

Variable2 = [1, 2, 2, 2, 3] → centrality = ? 2

Variable3 = [1, 200, 1000, 1000000] → centrality???

- Different measures: mean/median/mode

Central Tendency: Mean

Let x_1, x_2, \dots, x_N be a set of N values or observations.

- The **arithmetic mean** (or **average**) of this set of values is the sum of values of a data set divided by number of values

$$\text{mean}(x) = \bar{x} = \frac{\sum_{i=1}^N x_i}{N} = \frac{x_1 + x_2 + \dots + x_N}{N}$$

- Example

data = [5, 7, 19, 21]

mean = (5 + 7 + 19 + 21) / 4 = 13

Central Tendency: Mean

Sometimes, each value x_i in a data set may be associated with a weight w_i for $i = 1, 2, \dots, N$. The weights reflect the significance, importance, or occurrence frequency attached to their respective values. In this case, the **weighted arithmetic mean** or the **weighted average** is

$$\bar{X} = \frac{\sum_{i=1}^N w_i x_i}{\sum_{i=1}^N w_i} = \frac{w_1 x_1 + w_2 x_2 + \dots + w_N x_N}{w_1 + w_2 + \dots + w_N}$$

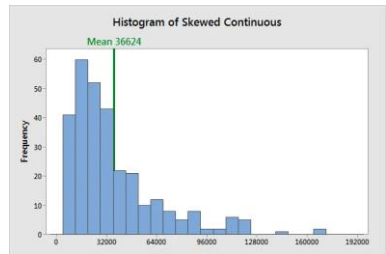
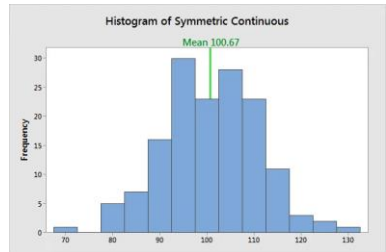
Central Tendency: Mean

- **Pro's:**

- Convenient summary
- Easy to understand

- **Con's:**

- Biased by:
 - extreme values
 - skewed distributions



Central Tendency: Median

- **Median:** value separating the **higher half** of a data sample, a population, or a probability distribution, **from the lower half**, i.e., “middle value”

$$\text{median}(x) = \begin{cases} x_{(r+1)} & \text{if } N \text{ is odd; i.e. } N=2r+1 \\ \frac{1}{2}(x_{(r)} + x_{(r+1)}) & \text{if } N \text{ is even; i.e. } N=2r \end{cases}$$

- Example

1, 3, 3, **6**, 7, 8, 9

Median = ?

1, 2, 3, **4**, **5**, 6, 8, 9

Median = ?

Central Tendency: Median

- **Pro's:**

- Robust to outliers and asymmetry

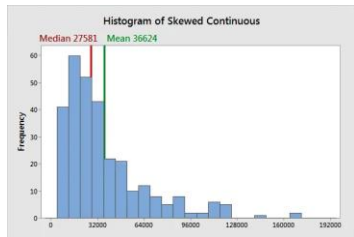
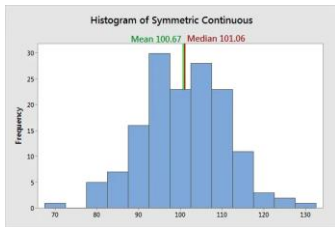
- **Con's:**

- Does not reflect asymmetry
e.g. 1, 1, 1, 1, 1, 1, 20 → median = 1

Median	Median Fixed
69	112
56	93
54	89
52	82
47	47
46	46
46	46
45	45
43	43
36	36
35	35
34	34
31	31

Median vs Mean

- Median and mean are equivalent for symmetrical distributions



Central Tendency: Mode

The **mode** for a set of data is the value that occurs most frequently in the set. Therefore, it can be determined for qualitative and quantitative attributes. It is possible for the greatest frequency to correspond to several different values, which results in more than one mode.

- **Example: value that occurs most often**

$$v_1 = [1, 1, 1, 3, 3, 5] \rightarrow ?$$

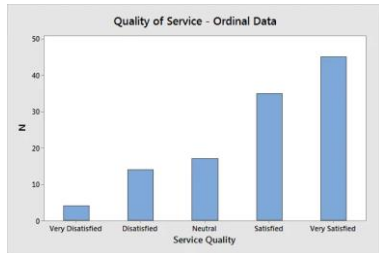
$$v_2 = [100, 200, 200, 1000000] \rightarrow ?$$

When there is a **tie**, the mode is generally **randomly selected**

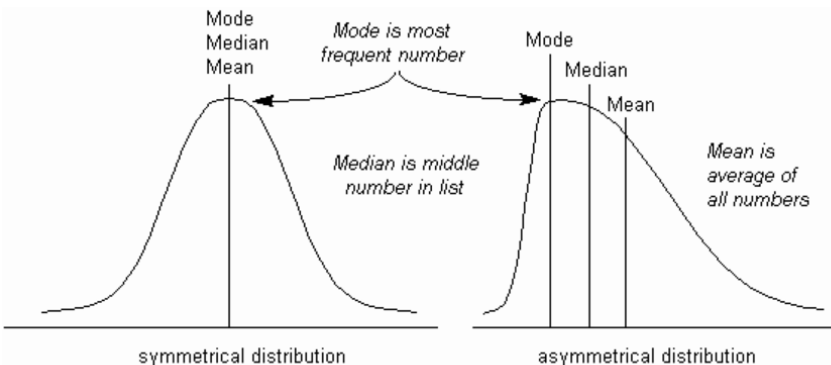
$$v_3 = [1, 1, 2, 2, 5, 7] \rightarrow ?$$

Central Tendency: Mode

- **Pro's:**
 - Good to get a sense of the most frequent value.
- **Con's:**
 - Is the mode always a central value?
 - e.g. data = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 10]



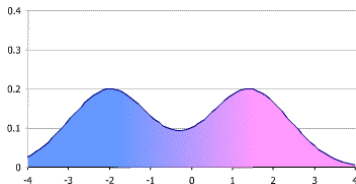
Mode, Median, and Mean



Unimodal vs Multimodal

Data sets with one mode are called *unimodal* and data sets with two or more modes are *multimodal*. At the other extreme, if each data value occurs only once, then there is no mode.

- Example of multimodal distribution



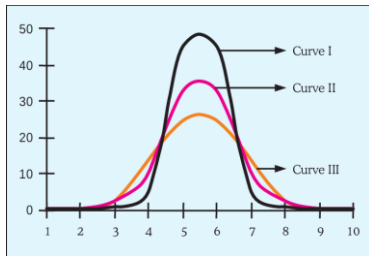
- Mode = one of the maxima.
- Mean = median = middle value.

WHICH ONE IS THE BEST?

- Symmetrical distribution
 - mean, median, and mode are equal
 - mean is preferred as it includes all the data
- Skewed distribution
 - Median is the best measure of central tendency
- Categorical data
 - Mode is preferred

Limitations of Using Only Central Tendency

- Many data distributions can have the same central value, e.g. mean.



DISPERSION (VARIABILITY, SPREAD)

range, interquartile range, variance, standard deviation

Dispersion of Data: Range

Let x_1, x_2, \dots, x_N be a set of observations for some numeric attribute, X . The **range** of the set is the difference between the largest ($\max(x)$) and smallest ($\min(x)$) values.

$$\text{range}(x) = \max(x) - \min(x) = x_{(N)} - x_{(1)}$$

Example: Difference between lowest value and highest value.

$$v_1 = [1, 1, 1, 3, 3, 5] \rightarrow ?$$

$$v_2 = [100, 200, 200, 1000000] \rightarrow ?$$

- Disadvantage: susceptible to outliers

Dispersion of Data: Interquartile range

Suppose that the data for attribute X is sorted in increasing numeric order. **Quantiles** are points taken at regular intervals of a data distribution, dividing it into essentially equal-size consecutive sets.

- **2-quantile** is the data point dividing the lower and upper halves of the data distribution, i.e. *median*
- **4-quantiles** are the three data points that split the data distribution into four equal parts, i.e. *quartiles*
- **100-quantiles** divide the data distribution into 100 equal-sized consecutive sets, i.e. *percentiles*

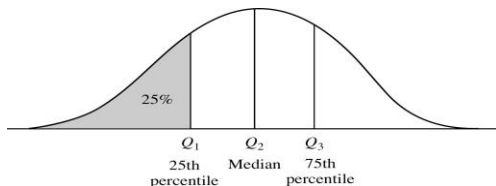
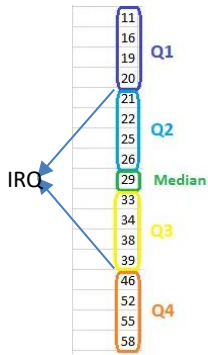
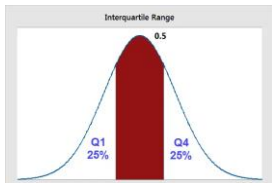


Figure: plotted percentiles for the data distribution of an attribute.

Dispersion of Data: Interquartile range

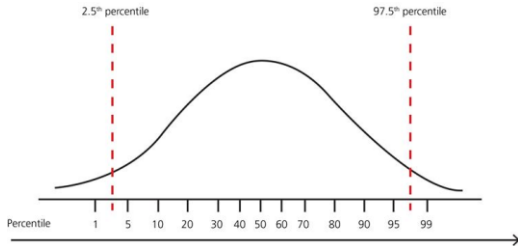
- Interquartile range: the middle half of the data
 - $Q3 - Q1$ (75 percentile - 25 percentile)
- Robust to outlier and skewed data



Dispersion of Data: Interquartile range

PERCENTILES

- Percentile: the value below which a percentage of data falls



Interquartile range = 75%-25%

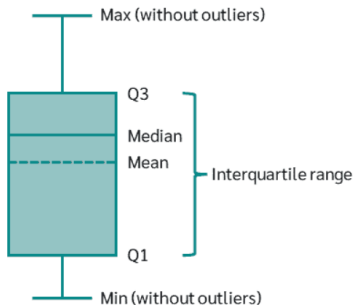
PERCENTILES

- Percentile: the value below which a percentage of data falls
- A list of data points: 2, 1, 3, 4, 5, 6, 7, 8, 9, 11, 10, 12
- Questions:
 - What is the 20-th percentile?
 - What is the percentile value for 6?

Dispersion of Data: Interquartile range

QUANTILES/PERCENTILE: DISCUSSION

- Can help determine if there is **symmetry**,
- Estimate the **dispersion**
- And find out if there are **extreme values**



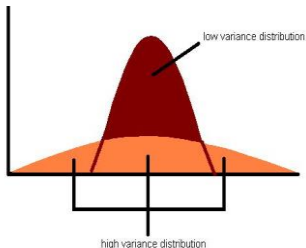
Dispersion of Data: variance

The variance of N observations, x_1, x_2, \dots, x_N , for a numeric attribute X is

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2$$

where \bar{x} is the mean value of the observations.

- Average squared difference of the values from the mean
- Variance measures how far a set of numbers is spread out
- Standard deviation: square root of variance



Dispersion of Data: variance

- Pro
 - A common measure to assess dispersion
- Con
 - Sensitive to large deviations around the mean, e.g., outliers

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2$$

WHICH IS THE BEST: STD, IQR, RANGE

- Normal distribution
 - STD with Mean
- Skewed distribution
 - IQR with Median
- Small sample size (not enough data)
 - Range

High Order Descriptive Statistics

- Mean:

$$\mu = \frac{\sum_{i=1}^n x_i}{n}$$

- Variance:

$$\text{var} = \sigma^2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n}$$

- Skewness:

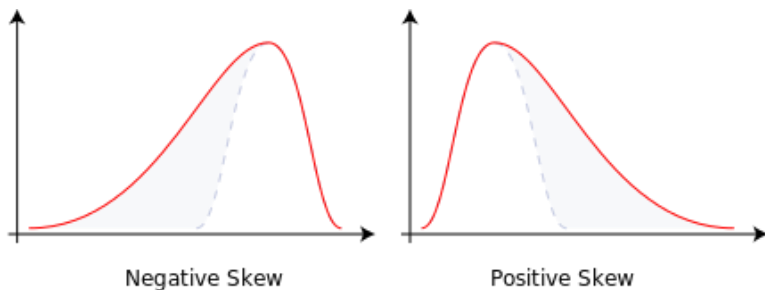
$$\gamma_1 = \frac{\sum_{i=1}^n (x_i - \mu)^3}{n\sigma^3}$$

- Kurtosis:

$$\gamma_2 = \frac{\sum_{i=1}^n (x_i - \mu)^4}{n\sigma^4}$$

SKEWNESS

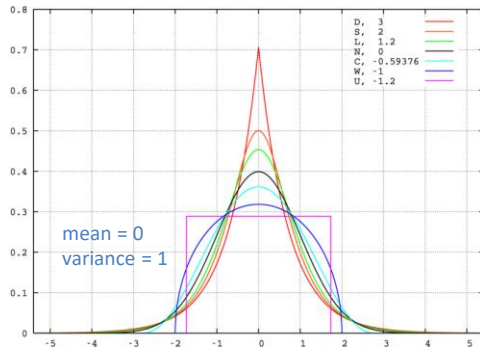
- Skewness captures asymmetry



High Order Descriptive Statistics

KURTOSIS

- Kurtosis captures “peakedness” (or concentration around the mean)



Python Implementation

```
import numpy as np
import scipy.stats as sts
x = np.random.normal(0,1,1000)

print([np.median(x), np.mean(x),
np.std(x), sts.skew(x), sts.kurtosis(x)])
print([np.percentile(x,q) for q in [0, 25, 50, 75, 100]])

hist,bins = np.histogram(x,100)
mode_idx = np.argmax(hist)
mode = bins[mode_idx]
```

Similarity and Distance

Similarity and Distance

Similarity and **dissimilarity measures** are referred to as measures of **proximity**. In data mining applications, such as *clustering*, *outlier analysis*, and *nearest-neighbor classification*, we need ways to assess how alike or unlike objects are in comparison to one another.

Data mining algorithms use the distance function as a key subroutine, and the design of the function directly impacts the quality of the results. Distance functions are highly sensitive to the type of the data, the dimensionality of the data, and the global and local nature of the data distribution.



Similarity and Dissimilarity Measures

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

Distance or dissimilarity between two objects is a numerical measure of the degree to which the two objects are different. Dissimilarities fall in the interval $[0, 1]$, but it is also common for them to range in the interval 0 to ∞ .

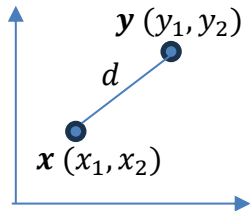
Euclidean Distance

The most popular distance measure is **Euclidean distance** (i.e. straight line). Let $\mathbf{x} = (x_1, x_2, \dots, x_p)$ and $\mathbf{y} = (y_1, y_2, \dots, y_p)$ be two objects described by p numeric attributes. The Euclidean distance between objects \mathbf{x} and \mathbf{y} is defined as:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_p - y_p)^2}$$

The Euclidean distance satisfies the following mathematical properties:

- **Non-negativity:** $d(\mathbf{x}, \mathbf{y}) \geq 0$
- **Identity of indiscernibles:** $d(\mathbf{x}, \mathbf{x}) = 0$
- **Symmetry:** $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$
- **Triangle inequality:** $d(\mathbf{x}, \mathbf{y}) \leq d(\mathbf{x}, \mathbf{k}) + d(\mathbf{k}, \mathbf{y})$



Minkowski Distance

The **Minkowski distance** is a generalization of the Euclidean distance. It is defined as

$$d(\mathbf{x}, \mathbf{y}) = \sqrt[h]{|x_1 - y_1|^h + |x_2 - y_2|^h + \cdots + |x_p - y_p|^h}$$

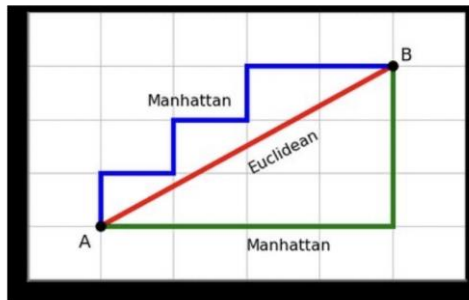
where h is a real number such that $h \geq 1$.

- **Euclidean distance**, when $h = 2$ (also known as L_2 norm)
- **Manhattan distance**, when $h = 1$ (i.e., L_1 norm).

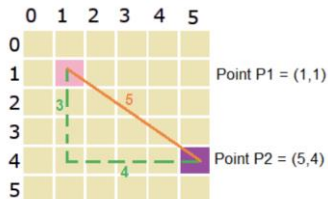
The Manhattan (or city block) distance is named so because it is the distance in blocks between any two points in a city

$$d(\mathbf{x}, \mathbf{y}) = |x_1 - y_1| + |x_2 - y_2| + \cdots + |x_p - y_p|$$

Euclidean and Manhattan Distances



omnicalculator.com



$$\text{Euclidean distance} = \sqrt{(5-1)^2 + (4-1)^2} = 5$$

$$\text{Manhattan distance} = |5-1| + |4-1| = 7$$

Cosine Similarity

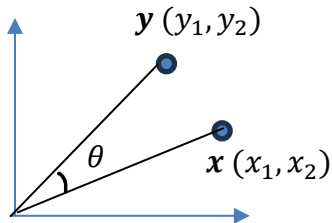
The **Cosine similarity** can be used to compare sparse vectors:

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

where $\|\mathbf{x}\|$ is the Euclidean norm of vector $\mathbf{x} = (x_1, x_2, \dots, x_p)$, defined as $\sqrt{x_1^2 + x_2^2 + \dots + x_p^2}$ and $\mathbf{x} \cdot \mathbf{y}$ is the inner product of \mathbf{x} and \mathbf{y} .

The measure computes the cosine of the angle between vectors \mathbf{x} and \mathbf{y} . A cosine value of 0 means that the two vectors have no match. The closer the cosine value to 1, the smaller the angle and the greater the match between vectors.

Cosine Similarity



$$\text{similarity} = \cos(\theta) = \frac{x \cdot y}{\|x\| \|y\|}$$

Cosine Similarity

Cosine similarity is often used to measure document similarity in text analysis (sparse data)

Document Vector or Term-Frequency Vector

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

Similarity Measures for Binary Data

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

- 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

	0	1	1	0	0	1	1	1
XOR	0	1	0	1	0	1	1	0
<hr/>								
	0	0	1	1	0	0	0	1

Simple Matching Coefficient

The **simple matching coefficient** (SMC) is defined as

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

This measure counts both presences and absences equally.

e.g. SMC can be used to find students who had answered questions similarly on a test that consisted only of true/false questions.

Jaccard Coefficient

The **Jaccard coefficient** is frequently used to handle objects consisting of asymmetric binary attributes. The Jaccard coefficient, which is often symbolized by J , is given by the following equation

$$J = \frac{\text{number of matching presences}}{\text{number of attributes not involved in 00 matches}} = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

Example: Calculate SMC and J for the following two binary vectors:

$x = (1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$ and

$y = (0, 0, 0, 0, 0, 0, 1, 0, 0, 1).$

Summary

Attributes and Objects Data sets are made up of data objects that are described by attributes.

Characteristics of Data Data sets have some well defined characteristics.

Data Representations Data representations are basic records, graphs, ordered.

Basic Statistical Descriptions of Data Basic statistical descriptions provide the analytical foundation for data preprocessing.

Similarity and Distance Different measures of proximity can be computed for specific attribute types.

Go back to menti.com and input the code: **2346 0479**
or

go this link:

<https://www.menti.com/aleq94qf4gzo>

