

EDA –Syllabus ISE Exam 2022

Introduction to data:

- Understanding data, Types of attributes, Nominal, ordinal, interval, ratio, Discrete and continuous attributes, Types of datasets: Record data, Graph-based data, Sequence data, time series data, spatial data, General characteristics of datasets
- Data quality problems, issues related to applications, Data transformations to make data suitable for data mining, EDA vs. classical data analytics

Exploring data using descriptive measures 12

- Frequency distribution : simple, grouped, cumulative and relative frequency distribution, graphs for frequency distribution (Histogram, frequency polygon, frequency curve, cumulative frequency curve)
 - Measures of central tendency: Mean (Arithmetic, weighted and geometric mean), , median, mode, mid range
 - Predicting missing data using regression modeling, interpolation
 - Measures of dispersion: range, inter-quartile range, variance, standard deviation, root mean square deviation, Coefficients of dispersion based upon range, quartile deviation, mean deviation, standard deviation, ANOVA.
 - Boxplot, Quantile–Quantile Plot, Scatter Plots and Data Correlation, Covariance, Bregman divergence.
 - Measures of Skewness: Pearson’s coefficient, Bowley’s coefficient, coefficient based upon moments
- **Data similarity and dissimilarity**
 - Similarity measures for numeric data, Minkowski distance, Euclidean distance, Manhattan distance, supremum distance, Mahalanobis distance, Bhattacharyya distance
 - Similarity measures for symmetric and asymmetric binary data, simple matching coefficient, Jaccard coefficient, hamming distance
 - Similarity measures for textual data, edit distance, cosine distance, Jaro distance, n-Gram distance , longest common subsequence, Dissimilarity between attributes of mixed type

Lecture 1: Data and Attributes

Mr. Vaibhav Chunekar

Assistant Professor

KJSCE

Date: 11/1/2022

What is a data?

- Data is a collection of facts. This collection can include numbers, pictures, videos, words, measurements, observations, and more.
- Data analysis is the collection, transformation, and organization of data in order to draw conclusions, make predictions, and drive informed decision-making.
- Data analytics can give us new information throughout data's entire life cycle.
-

How Data Analysis helps?

- Data is everywhere. You use and create data everyday. Have you ever read reviews of a product before deciding whether or not to buy it? That's data analysis.
- Or maybe you wear a fitness tracker to count your steps so you can stay active throughout the day. That's data analysis

How do you create Data and role of Data Analyst?

- You also create huge amounts of it every single day.
- Any time you use your phone, look up something online, stream music, shop with a credit card, post on social media or use GPS to map a route, you're creating data.
- Data analyst is someone who collects, transforms, and organizes data in order to help make informed decisions

What is Data?

- Collection of ***data objects*** and their ***attributes***
- An ***attribute*** is a property or characteristic of an object
 - Examples: eye color of a person, temperature, etc.
 - Attribute is also known as variable, field, characteristic, dimension, or feature
- A collection of attributes describe an ***object***
 - Object is also known as record, point, case, sample, entity, or instance

Attributes

Objects

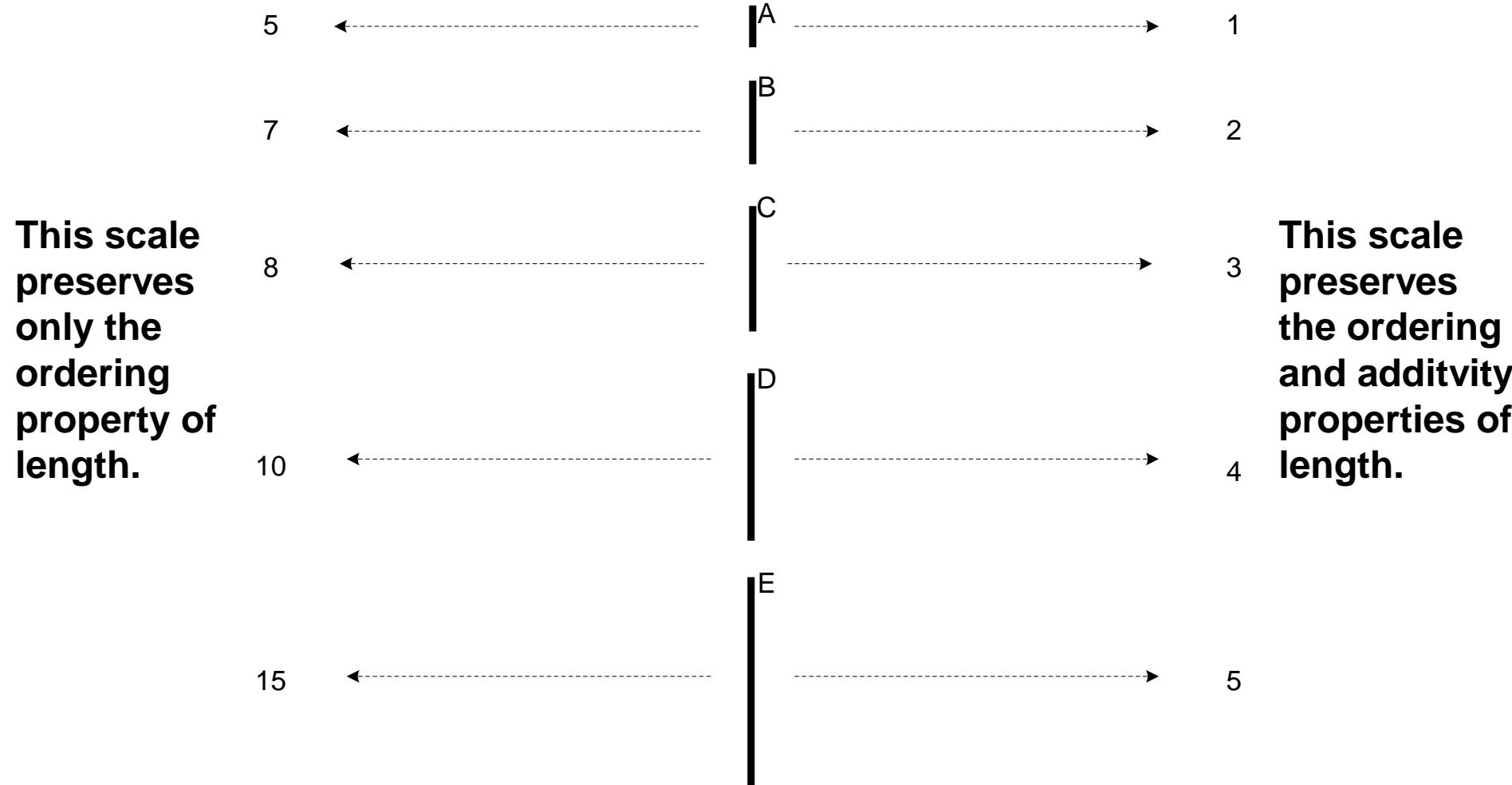
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Attribute Values

- **Attribute values** are numbers or symbols assigned to an attribute for a particular object
- 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 can be different than the properties of the values used to represent the attribute

Measurement of Length

- The way you measure an attribute may not match the attribute's properties.



Types of Attributes

- There are different types of attributes
 - Nominal
 - Examples: ID numbers, eye color, zip codes
 - Ordinal
 - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height {tall, medium, short}
 - Interval
 - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
 - Ratio
 - Examples: temperature in Kelvin, length, counts, elapsed time (e.g., time to run a race)

Properties of Attribute Values

- The type of an attribute depends on which of the following properties/operations it possesses:
 - Distinctness: $= \neq$
 - Order: $< >$
 - Differences are $+$ $-$
meaningful :
 - Ratios are $*$ $/$
meaningful
 - Nominal attribute: distinctness
 - Ordinal attribute: distinctness & order
 - Interval attribute: distinctness, order & meaningful differences
 - Ratio attribute: all 4 properties/operations

Difference Between Ratio and Interval

- Is it physically meaningful to say that a temperature of 10° is twice that of 5° on
 - the Celsius scale?
 - the Fahrenheit scale?
 - the Kelvin scale?
- Consider measuring the height above average
 - If Bill's height is three inches above average and Bob's height is six inches above average, then would we say that Bob is twice as tall as Bill?
 - Is this situation analogous to that of temperature?

	Attribute Type	Description	Examples	Operations
Categorical Qualitative	Nominal	Nominal attribute values only distinguish. ($=$, \neq)	zip codes, employee ID numbers, eye color, sex: { <i>male</i> , <i>female</i> }	mode, entropy, contingency correlation, χ^2 test
	Ordinal	Ordinal attribute values also order objects. ($<$, $>$)	hardness of minerals, { <i>good</i> , <i>better</i> , <i>best</i> }, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Numeric Quantitative	Interval	For interval attributes, differences between values are meaningful. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, <i>t</i> and <i>F</i> tests
	Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, current	geometric mean, harmonic mean, percent variation

Attribute Type	Transformation	Comments
Categorical Qualitative	Nominal	Any permutation of values If all employee ID numbers were reassigned, would it make any difference?
	Ordinal	An order preserving change of values, i.e., $new_value = f(old_value)$ where f is a monotonic function An attribute encompassing the notion of good, better best can be represented equally well by the values {1, 2, 3} or by { 0.5, 1, 10}.
Numeric Quantitative	Interval	$new_value = a * old_value + b$ where a and b are constants Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree).
	Ratio	$new_value = a * old_value$ Length can be measured in meters or feet.

Summary

- What is a Data? Role of Data Analyst in a society?
- Exploratory Data Analysis- Data and Attributes
- Measurements of Length
- Attributes and Types: categorical(Qualitative) and Numerical(Quantitative)
- Categorical Types: Nominal , Ordinal ,
- Quantitative Types: Interval and Ratio(Example and operation)
- Examples with Transformation and operation possible on each attributes

Lecture 2: Data and Dataset

Date: 12/1/2022

Data Objects

- Data sets are made up of data objects.
- A **data object** represents an entity.
- Examples:
 - sales database: customers, store items, sales
 - medical database: patients, treatments
 - university database: students, professors, courses
- Also called *samples*, *examples*, *instances*, *data points*, *objects*, *tuples*.
- Data objects are described by **attributes**.
- Database rows -> data objects; 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*
- Types:
 - Nominal
 - Binary
 - Ordinal
 - Numeric: quantitative
 - Interval-scaled
 - Ratio-scaled

Attribute Types

- **Nominal:** categories, states, or “names of things”, do not have any meaningful order, enumeration
 - $Hair_color = \{auburn, black, blond, brown, grey, red, white\}$
 - marital status, occupation = Not =
- **Binary**
 - Nominal attribute with only 2 states (0 and 1)
 - Symmetric binary: both outcomes equally important, have same weight
 - e.g., gender
 - Asymmetric binary: outcomes not equally important.
 - e.g., medical test (positive vs. negative)
 - Convention: assign 1 to most important outcome (e.g., HIV positive)
- **Ordinal**
 - Values have a meaningful order (ranking) but magnitude between successive values is not known. < >
 - $Size = \{small, medium, large\}$, grades{A,B,C,D}, army rankings

Numeric Attribute Types

- Quantity (integer or real-valued)
- **Interval**
 - Measured on a scale of **equal-sized units**
 - Values have order + -
 - E.g., *temperature in C° or F°, calendar dates*
 - No true zero-point
 - Can compare and quantify
 - Numeric in nature
 - Measures of central tendency(mean, median, mode)
- **Ratio**
 - Inherent **zero-point**
 - We can speak of values as being an order of magnitude larger than the unit of measurement (10 K° is twice as high as 5 K°).
 - e.g., *temperature in Kelvin, length, counts, monetary quantities*

Discrete vs. Continuous Attributes

- **Attributes ML point of view**
- **Discrete Attribute**
 - Has only a finite or countably infinite (one to one correspondence with natural number) set of values
 - E.g., 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
- **Continuous Attribute**
 - Has real numbers as attribute values
 - E.g., 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

Asymmetric Attributes

- Only presence (a non-zero attribute value) is regarded as important
 - Words present in documents
 - Items present in customer transactions
- If we met a friend in the grocery store would we ever say the following?

“I see our purchases are very similar since we didn’t buy most of the same things.”

Critiques of the attribute categorization

- Incomplete
 - Asymmetric binary
 - Cyclical
 - Multivariate
 - Partially ordered
 - Partial membership
 - Relationships between the data
- Real data is approximate and noisy
 - This can complicate recognition of the proper attribute type
 - Treating one attribute type as another may be approximately correct

Key Messages for Attribute Types

- The types of operations you choose should be “meaningful” for the type of data you have
 - Distinctness, order, meaningful intervals, and meaningful ratios are only four (among many possible) properties of data
 - The data type you see – often numbers or strings – may not capture all the properties or may suggest properties that are not present
 - Analysis may depend on these other properties of the data
 - Many statistical analyses depend only on the distribution
 - In the end, what is meaningful can be specific to domain

Important Characteristics of Data

- Dimensionality (number of attributes)
 - High dimensional data brings a number of challenges
- Sparsity
 - Only presence counts.
 - Skew data
 - Non zero values are important
- Resolution
 - Patterns depend on the scale
- Size
 - Type of analysis may depend on size of data

Summary

- Revise the Attributes and types
- Discrete and Continuous attributes
- Asymmetric data attributes
- Guidelines or comments on attributes of data
- Characteristics of data

Lecture 3: Types of Dataset

Date: 13/1/2022

Review General characteristics of Data Sets

- **Dimensionality, Sparsity, and Resolution.**

Dimensionality of a data

- The dimensionality of a data set is the number of attributes that the objects in the data set have.
- In a particular data set if there are high number of attributes (also called high dimensionality), then it can become difficult to analyse such a data set. When this problem is faced, it is referred to as **Curse of Dimensionality**.
- In order to understand what the hell is this **Curse of Dimensionality**, we first need to understand the other two characteristics of Data.

Sparsity

- For some data sets, such as those with asymmetric features, most attributes of an object have values of 0; in many cases fewer than 1% of the entries are non-zero. Such a data is called **sparse data** or it can be said that the data set has **Sparsity**.

Resolution

- The patterns in the data depend on the level of resolution. If the resolution is too fine, a pattern may not be visible or may be buried in noise; if the resolution is too coarse, the pattern may disappear. For example, variations in atmospheric pressure on a scale of hours reflect the movement of storms and other weather systems. On a scale of months, such phenomena are not detectable

Curse of Dimensionality

- Many types of Data Analysis becomes difficult as the dimensionality (number of attributes in the data set) of the data set increases.
- Specifically, as dimensionality increases, the data becomes increasingly sparse in the space that it occupies.
- For classification, this can mean that there are not enough data objects to allow the creation of a model that reliably assigns a class to all possible objects.
- For clustering, the definitions of density and the distance between points, which are critical for clustering, become less meaningful.

Data sets

- A Data set is a set or collection of data.
- This set is normally presented in a tabular pattern.
- Every column describes a particular variable.
- And each row corresponds to a given member of the data set, as per the given question.
- This is a part of data management

Types of Data Sets in statistics

- Numerical data sets
- Bivariate data sets
- Multivariate data sets
- Categorical data sets
- Correlation data sets

Numerical Data Sets

- The numerical data set is a data set, where the data are expressed in numbers rather than natural language.
- The numerical data is sometimes called quantitative data. The set of all the quantitative data/numerical data is called the numerical data set.
- The numerical data is always in the numbers form, such that we can perform arithmetic operations on it.
- Weight and height of a person
- The count of RBC in a medical report
- Number of pages present in a book

Bivariate Data Sets

- A data set that has two variables is called a Bivariate data set. It deals with the relationship between the two variables. Bivariate dataset usually contains two types of related data.
- Example: To find the percentage score and age of the students in a class. Score and age can be considered as two variables
- The sales of ice cream versus the temperature on that day. Here the two variables used are ice cream and temperature.
- (Note: In case, if you have one set of data alone say, temperature, then it is called the univariate dataset)

Multivariate Data Sets

- A data set with multiple variables. When the dataset contains three or more than three data types (variables), then the data set is called a multivariate dataset. In other words, the multivariate dataset consists of individual measurements that are acquired as a function of three or more than three variables.
- Example: If we have to measure the length, width, height, volume of a rectangular box, we have to use multiple variables to distinguish between those entities.

Categorical Data Sets

- Categorical data sets represent features or characteristics of a person or an object. The categorical dataset consists of a categorical variable also called the qualitative variable, that can take exactly two values. Hence, it is termed as a dichotomous variable. Categorical data/variables with more than two possible values are called polytomous variables. The qualitative/categorical variables are often assumed to be polytomous variable unless otherwise specified.
- Example:
- A person's gender (male or female)
- Marital status (married/unmarried)

Correlation Data Sets

- The set of values that demonstrate some relationship with each other indicates correlation data sets. Here the values are found to be dependent on each other.
- Generally, correlation is defined as a statistical relationship between two entities/variables. In some scenarios, you might have to predict the correlation between the things. It is essential to understand how correlation works. The correlation is classified into three types. They are:
- Positive correlation – Two variables move in the same direction (Either both are up or both or down)
- Negative correlation – Two variables move in opposite directions. (One variable is up and another variable is down and vice versa)
- No or zero correlation – No relationship between two variables.
- Example: A tall person is considered to be heavier than a short person. So here the weight and height variables are dependent on each other

Types of data sets in Data Mining Domain

- Record
 - Data Matrix
 - Document Data
 - Transaction Data
- Graph
 - World Wide Web
 - Molecular Structures
- Ordered
 - Spatial Data
 - Temporal Data
 - Sequential Data
 - Genetic Sequence Data

Record Data

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

Tid	Refund	Marital Status	Taxable Income	Defaulted Borrower
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

(a) Record data.

TID	ITEMS
1	Bread, Soda, Milk
2	Beer, Bread
3	Beer, Soda, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Soda, Diaper, Milk

(b) Transaction data.

Projection of x Load	Projection of y Load	Distance	Load	Thickness
10.23	5.27	15.22	27	1.2
12.65	6.25	16.22	22	1.1
13.54	7.23	17.34	23	1.2
14.27	8.43	18.45	25	0.9

(c) Data matrix.

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

(d) Document-term matrix.

Record Data

- Transaction data- Record data → Market basket data
- Market basket data because the items in each record are the products in persons market basket.

Transaction or Market Basket Data

- It is a special type of record data, in which each record contains a set of items. For example, shopping in a supermarket or a grocery store. For any particular customer, a record will contain a set of items purchased by the customer in that respective visit to the supermarket or the grocery store. This type of data is called **Market Basket Data**. Transaction data is a collection of sets of items, but it can be viewed as a set of records whose fields are asymmetric attributes. Most often, the attributes are binary, indicating whether or not an item was purchased or not.

Data Matrix (pattern 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 a 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

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 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.

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

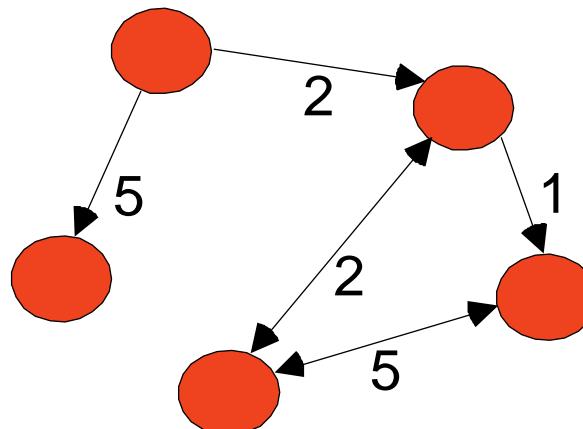
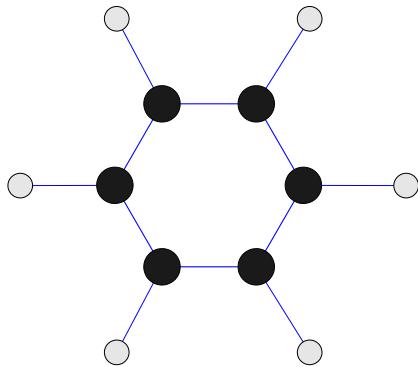
Transaction Data

- A special type of data, where
 - Each transaction involves a set of items.
 - For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.
 - Can represent transaction data as record data

<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

Graph Data : Molecular(Data with Objects)

- Examples: Generic graph, a molecule, and webpages : Graph Mining : Analyze Graph Data



Benzene Molecule: C_6H_6

WWW dataset:

- Text and link other page
- Search Queries on WWW: we collect web pages and process web pages and extract data.

Useful Links:

- [Bibliography](#)
- Other Useful Web sites
 - [ACM SIGKDD](#)
 - [KDnuggets](#)
 - [The Data Mine](#)

Knowledge Discovery and Data Mining Bibliography

(Gets updated frequently, so visit often!)

- [Books](#)
- [General Data Mining](#)

Book References in Data Mining and Knowledge Discovery

Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Ithurusamy, "Advances in Knowledge Discovery and Data Mining", AAAI Press/the MIT Press, 1996.

J. Ross Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, 1993.
Michael Berry and Gordon Linoff, "Data Mining Techniques (For Marketing, Sales, and Customer Support)", John Wiley & Sons, 1997.

General Data Mining

Usama Fayyad, "Mining Databases: Towards Algorithms for Knowledge Discovery", Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, vol. 21, no. 1, March 1998.

Christopher Matheus, Philip Chan, and Gregory Piatetsky-Shapiro, "Systems for knowledge Discovery in databases", IEEE Transactions on Knowledge and Data Engineering, 5(6):903-913, December 1993.

Summarized

- Revise Characteristics of data and watch imp. Curse of Dimensionality
- Data Set meaning
- Exploring Statistical types of datasets: univariate ,multivariate...
- Data Mining Datasets:
 - 1. Record Data :Flat File(RDBMS), Transaction Data(Market Basket Transaction Data),Data Matrix(Pattern Matrix), Document Data (e.g. Marksheets of students),sparse Matrix
 - 2. Graph Dataset: WWW dataset : Data and link
Molecular Dataset: Relationship among object(Web Mining)

Thank you!!!

Lecture 4

Ordered Data and Data Quality

17/1/2022

Last week Revision

- Data:
- Attributes of Data-Qualitative(Categorical)and Quantitative(Numeric)
- Datasets- Dimensionality-curse of dimensionality-sparsity , resolution
- Types of data sets-Record Data(Transaction ,document, sparse,Matrix)
- Graph based data-WWW and molecular data

Ordered Data

- For some types of data, the attributes have relationships that involve order in time or space. This data can be segregated into *four* types:

1. Sequential Data

2. Sequence Data

3. Time Series Data

4. Spatial Data

Sequential Data(Sequential Transaction Data)

- Also referred to as **temporal** data.
- It can be thought of as an extension of record data,
- where each record has a time associated with it.
- Example :

Consider a retail transaction data set that also stores the time at which the transaction took place

Time	Customer	Items Purchased
t1	C1	A, B
t2	C3	A, C
t2	C1	C, D
t3	C2	A, D
t4	C2	E
t5	C1	A, E

Customer	Time and Items Purchased
C1	(t1: A,B) (t2:C,D) (t5:A,E)
C2	(t3: A, D) (t4: E)
C3	(t2: A, C)

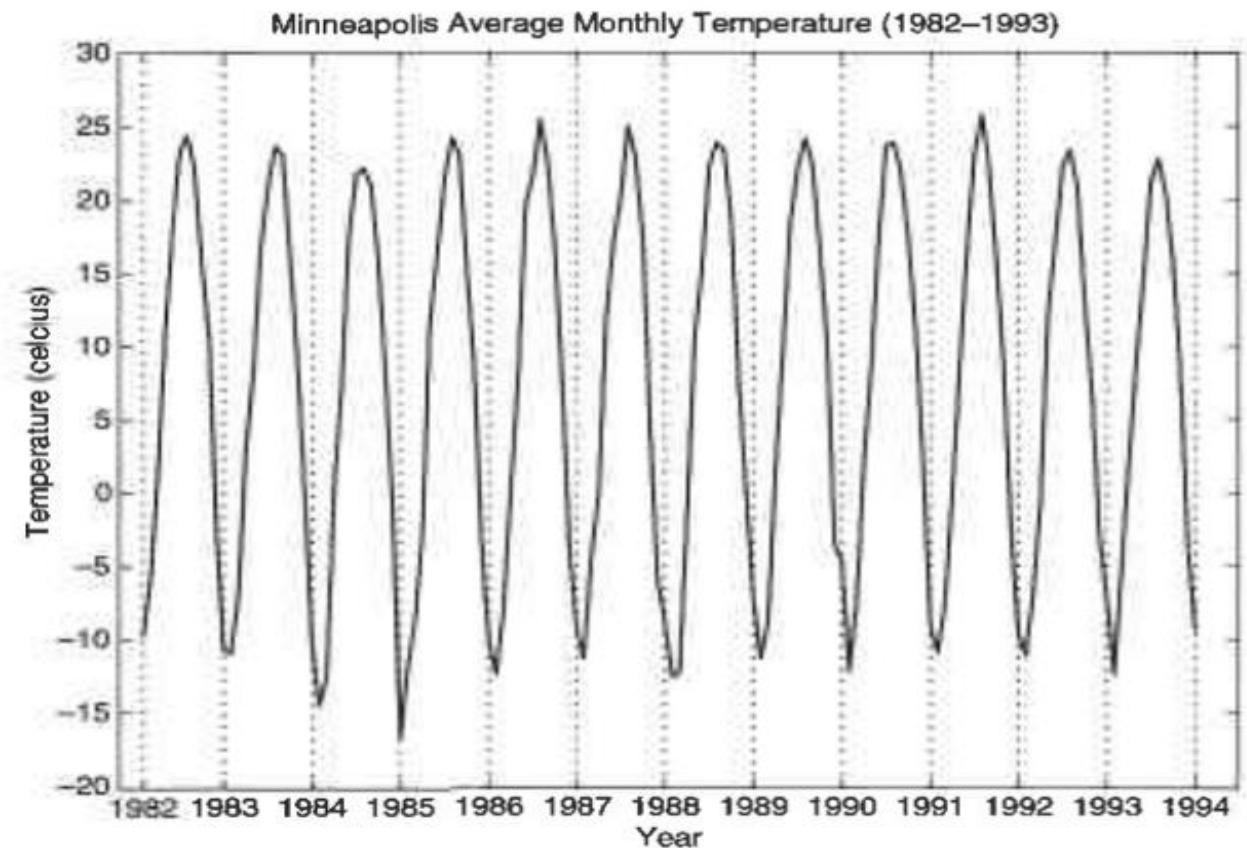
Sequence Data(e.g. Genomic Sequence data)

- Sequence data consists of a data set that is a sequence of individual entities, such as a sequence of words or letters.
- It is quite similar to sequential data, except that there are no time stamps; instead, there are positions in an ordered sequence.
- For example, the genetic information of plants and animals can be represented in the form of sequences of nucleotides that are known as genes.

GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCC GCCCGCGCCGTC
GAGAAGGGCCC GCCTGGCGGGCG
GGGGGAGGC GGGGCCGCCC GAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGC GGGCAGCGGACAG
GCCAAGTAGAACAC GCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG

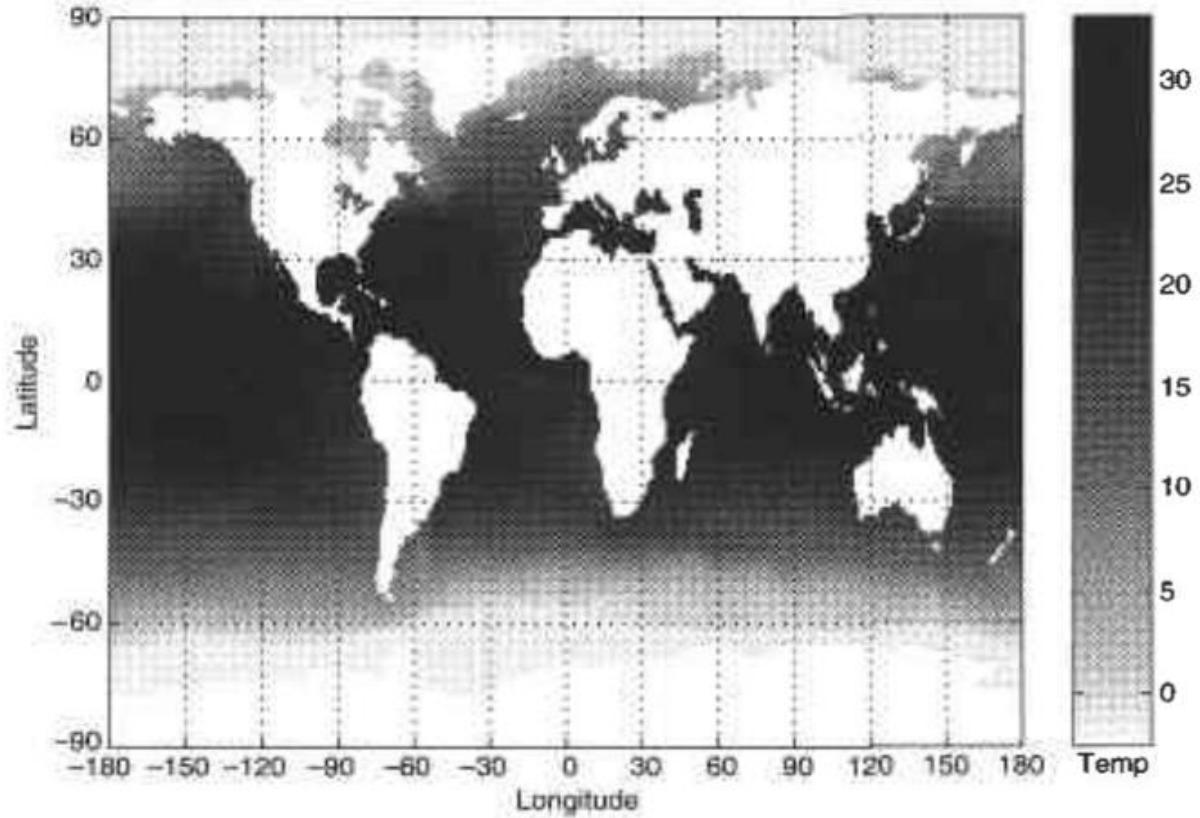
Time Series Data

- It's a special type of sequential data .
- Record is a time series, i.e., a series of measurements taken over time.
- For example: a financial data set might contain objects that are time series of the daily prices of various stocks.



Spatial Data

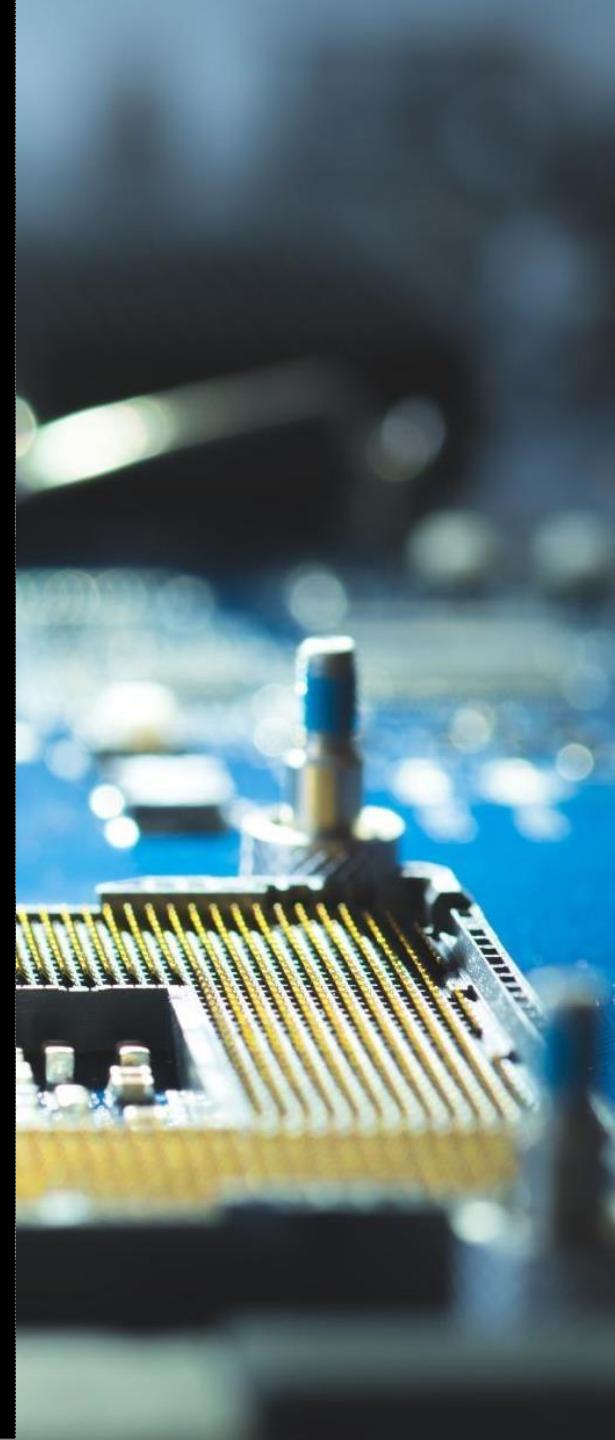
- Some objects have spatial attributes, such as positions or areas, as well as other types of attributes.
- An example of spatial data is weather data (precipitation, temperature, pressure) that is collected for a variety of geographical locations.



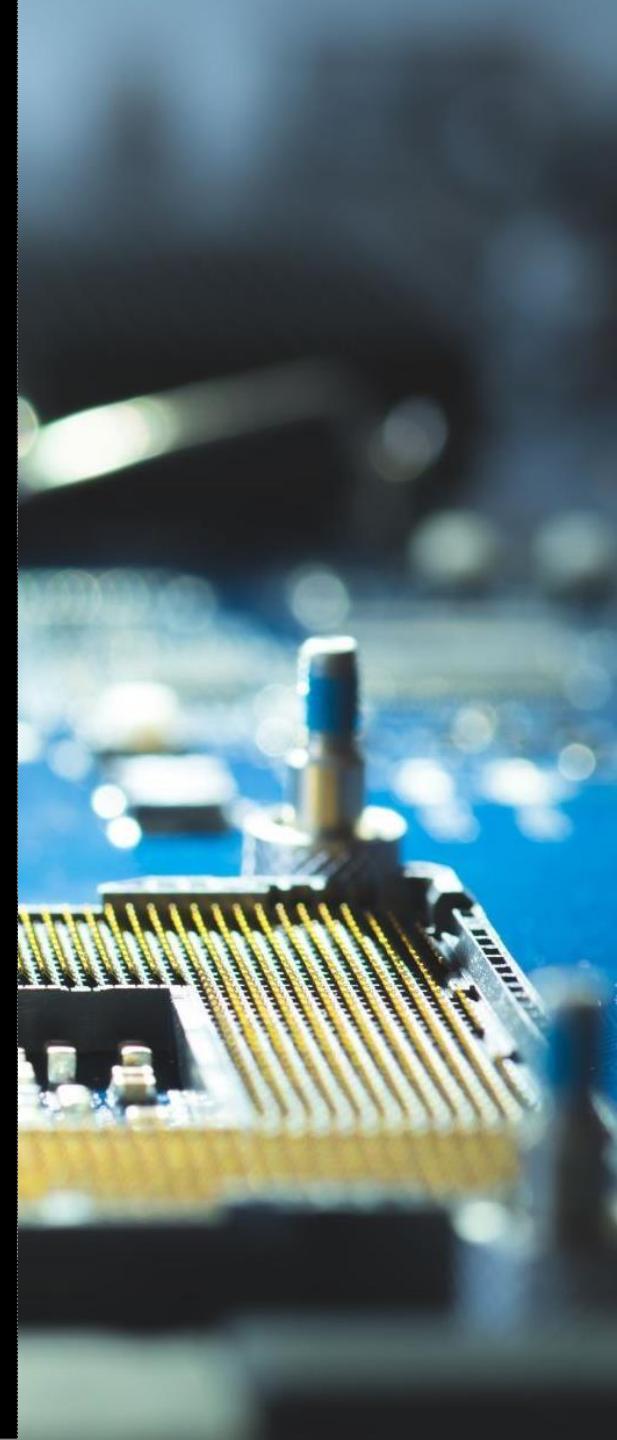
Non Record Based Data

- Example Benzene Ring-Substructure of atoms/molecules
- Not Capture all the information in the data
- Such data –data matrix where rows-locations and columns time

Data Quality



Data Quality: Why preprocess the data?





Data Cleaning

- Data mining focuses on (1) the detection and correction of data quality problems and (2) the use of algorithms that can tolerate poor data quality.
- **Data cleaning (or data cleansing) routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.**





Data Quality Problems

- **Low-quality data will lead to low-quality mining.**
Data is susceptible to noise, data missing and inconsistency.
 - **Example:** A classification model for detecting people who are loan risks is built using poor data
 - Some credit-worthy candidates are denied loans
 - More loans are given to individuals that default
- ## CONCERNS
- What are the **elements defining data quality?**
 - What **kinds of data quality problems?**
 - How can we **detect problems with the data?**
 - What can we **do about these problems?**
 - How can the **data be preprocessed** in order to help **improve the quality of the data** and, consequently, improve **efficiency of mining results** and ease it?

Elements defining Data quality problems

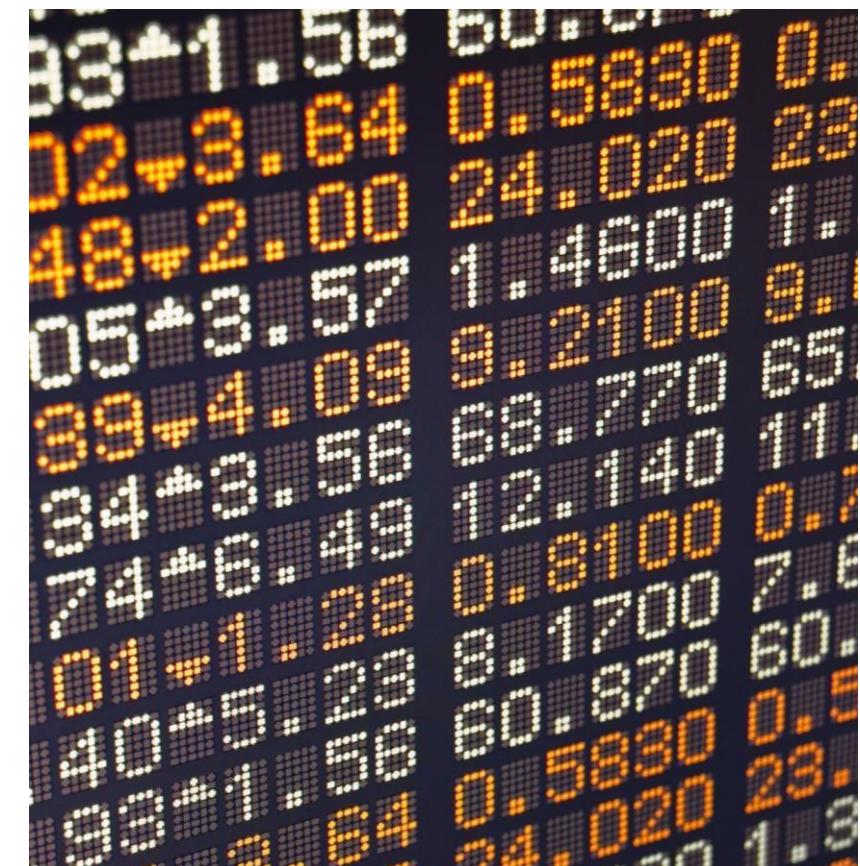
- The elements defining data quality: **accuracy, completeness, consistency, timeliness, believability and interpretability.**
- **Incompleteness:**
 - Attributes of interest may not always be available, such as customer information for sales transaction data.
 - Data may not be included simply because they were not considered important at the time of entry.
 - Relevant data may not be recorded due to a misunderstanding or because of equipment malfunctions.
 - Data that were inconsistent with other recorded data may have been deleted.
- **Consistency:** some modified but some not, dangling
- **Timeliness:** timely update
- **Believability:** how trustable the data are correct?
- **Interpretability:** how easily the data can be understood?

Measurement and Data Collection Issues

- **Inaccurate Data due to measurement and data collection errors:**
 - Data collection instruments used may be **faulty**.
 - Human or computer errors occurring at **data entry**.
 - Values or even entire data objects may be **missing**.
 - Users may purposely submit **incorrect data values** for mandatory fields when they do not wish to submit personal information(**disguised missing data**.)
 - Technology limitations such as **limited buffer size for coordinating synchronized data transfer and consumption**.
 - **Inconsistencies in naming conventions** or data codes, or inconsistent formats for input fields (e.g date).
 - **Duplicate tuples**

Data Collection Errors & Measurement Errors

- **Error** refers to the numerical difference of the measured and true value for the attributes.
 - **Data Collection Error** refers to errors such as omitting data objects or attribute values, or inappropriately including a data object.
 - **Measurement Error** refers to any problem resulting from the measurement process ie the value recorded differs from the true value to some extent.
 - The **quality of the measurement process** and the resulting data are measured by **precision, bias and accuracy**.
 - **Precision:** The closeness of repeated measurements (of the same quantity) to one another. Precision is often measured by the **standard deviation of a set of values**.
 - **Bias:** A systematic variation of measurements from the quantity being measured. Bias is measured by taking the **difference between the mean of the set of values and the known value of the quantity being measured**.
 - **Accuracy:** The closeness of measurements to the true value of the quantity being measured.





EXAMPLE

- We have a standard laboratory weight with a mass of **1g** and want to assess the precision and bias of our new laboratory scale. We weigh the mass five times, and obtain the following five values: **{1.015, 0.990, 1.013, 1.001, 0.986}**. Calculate the **bias** and **precision**.

Data Quality Summary

- It refers to the ability of a data set to serve whichever need a company hopes to use it for.
- That need
 - could be sending marketing materials to customers.
 - could be studying the market to plan a new product feature.
 - Could be maintaining a database of customer data for help with product support services
 - Could any number of other goals.
- No matter what the exact use case for your data, data quality is important.
- Without it, the data can't fulfill its intended purpose.
- Errors within a database of addresses would prevent you from using the data to reach customers effectively.
- A database of phone numbers that doesn't always include area codes for each entry falls short of providing the information you need to put the data to use in many situations.

Common causes of data quality problems

- Manual Entry
- OCR errors
- Lack of complete information
- Ambiguous data
- Duplicate data
- Data transformation errors

Common Errors

- 1. Manual data entry errors
- Humans are prone to making errors, and even a small data set that includes data entered manually by humans is likely to contain mistakes. Data entry errors such as typos, data entered in the wrong field, missed entries, and so on are virtually inevitable.
- 2. OCR errors
- Machines can make mistakes when entering data, too. In cases where organizations must digitize large amounts of data quickly, they often rely on Optical Character Recognition, or OCR, technology to do so. OCR technology scans images and extracts text from them automatically. It can be very useful when, for example, you want to take thousands of addresses that are printed on paper and enter them into a digital database so you can analyze them. The problem with OCR is that it is almost always imperfect.
- If you're OCR'ing thousands of lines of text, you're almost certainly going to have some characters or words that are misinterpreted – zeroes that are interpreted as eights, for example, or proper nouns that are read as common words because the OCR tool fails to distinguish properly between capital and lowercase letters. The same sorts of issues arise with other types of automated machine entry of data, such as text-to-speech

Common Errors

3. Lack of complete information

- When compiling a data set, you frequently run into the problem of not having all information available for every entry.
- For example, a database of addresses may be missing the zip codes for some entries because the [zip codes](#) couldn't be determined via the method that was used to compile the dataset.

4. Ambiguous data

- When building a database, you may find that some of your data is ambiguous, leading to uncertainty about whether, how and where to enter it.
- For example, if you are creating a database of phone numbers, some of the numbers you seek to enter may be longer than the typical ten digits that you have in a United States phone number. Are those longer numbers simply typos, or are they international phone numbers that include more digits? In the latter case, does the number contain complete international dialing information?
- These are the sorts of questions that are hard to answer quickly and systematically when you're working with a large body of data.

Common Error

5. Duplicate data

- You may find that two or more data entries are mostly or completely identical.
- For example, maybe your database contains two entries for a John Smith living at 123 Main St. Based on this information, it's difficult to know whether these entries are simply duplicates (maybe John Smith's information was entered twice by mistake) or if there are two John Smiths (a father and son, perhaps) living at the same address. You need to sort out seemingly duplicate entries like this to make the best use of your data.

6. Data transformation errors

- Converting data from one format to another can lead to mistakes.
- As a simple example, you may have a spreadsheet that you convert to a comma-separated value, or CSV file. Because data fields inside CSV files are separated by commas, you may run into issues when performing this conversion in the event that some of the data entries in your spreadsheet contain commas inside them.
- Unless your data conversion tools are sufficiently smart, they won't know the difference between a comma that is supposed to separate two data fields and one that is an internal part of a data entry. This is a basic example; things get much more complicated when you must perform complex data conversions, such as taking a mainframe database that was designed decades ago and converting it to NoSQL, a category of database that has become popular in just the last few years.

Problems Data Quality

- Poor data quality negatively affects many data processing efforts
- Data mining example: a classification model for detecting people who are loan risks is built using poor data
 - Some credit-worthy candidates are denied loans
 - More loans are given to individuals that default

Data Quality Problems ...

- What kinds of data quality problems?
- How can we detect problems with the data?
- What can we do about these problems?
- Examples of data quality problems:
 - Noise and outliers
 - Wrong data
 - Fake data
 - Missing values
 - Duplicate data

TYPES OF DATA QUALITY PROBLEMS

Data in the Real World Is Dirty. Data has to be cleaned due to:



Noise and outliers



Missing values or Incomplete Data



Duplicate data



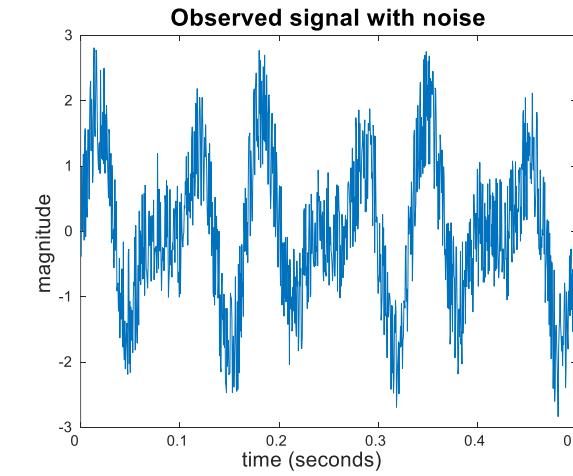
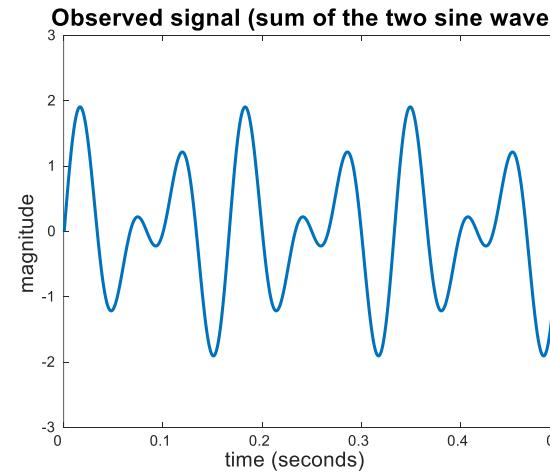
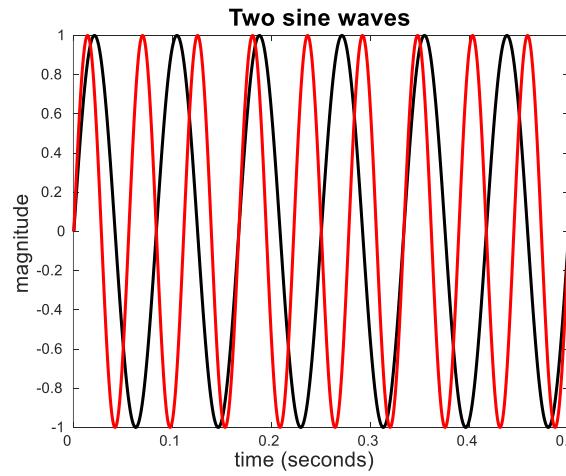
Wrong data or Inconsistent Data



Fake data or Disguised Missing Data

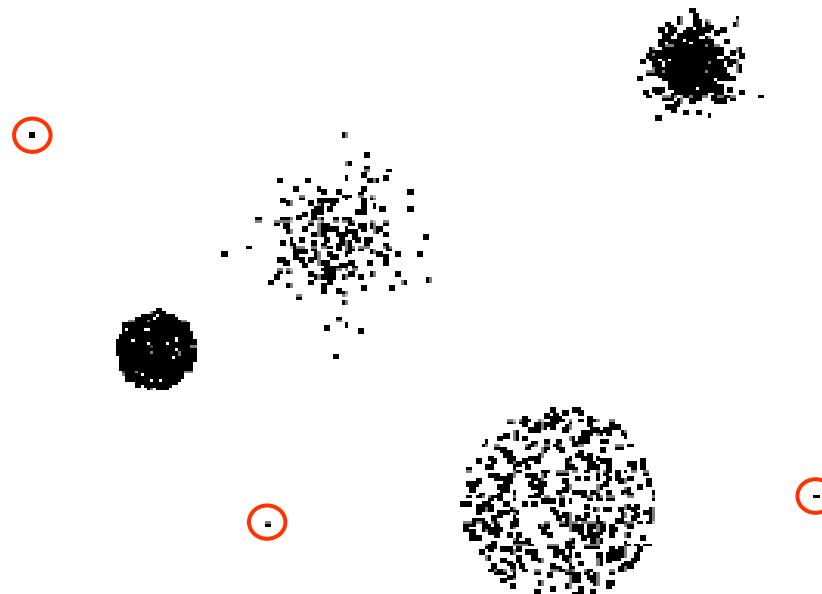
Noise

- For objects, noise is an extraneous object
- For attributes, noise refers to modification of original values
 - Examples: distortion of a person's voice when talking on a poor phone and "snow" on television screen
 - The figures below show two sine waves of the same magnitude and different frequencies, the waves combined, and the two sine waves with random noise
 - The magnitude and shape of the original signal is distorted



Outliers

- **Outliers** are data objects with characteristics that are considerably different than most of the other data objects in the data set
 - **Case 1:** Outliers are noise that interferes with data analysis
 - **Case 2:** Outliers are the goal of our analysis
 - Credit card fraud
 - Intrusion detection
- Causes?



Missing Values

- Reasons for missing values
 - Information is not collected
(e.g., people decline to give their age and weight)
 - Attributes may not be applicable to all cases
(e.g., annual income is not applicable to children)
- Handling missing values
 - Eliminate data objects or variables
 - Estimate missing values
 - Example: time series of temperature
 - Example: census results
 - Ignore the missing value during analysis

Duplicate Data

- Data set may include data objects that are duplicates, or almost duplicates of one another
 - Major issue when merging data from heterogeneous sources
- Examples:
 - Same person with multiple email addresses
- Data cleaning
 - Process of dealing with duplicate data issues
- When should duplicate data not be removed?

Summary

- Ordered Data
- Data Quality
- Causes of Data Quality?
- What is the problem with data Quality?

Lecture 5

Issues Related to Application in Data Quality

Vaibhav Chunekar

19/1/2022

Revision

- Ordered Data
- Data Quality:
 - Poor data quality negatively affects many data processing efforts
- Causes of Data Quality:
 - Manual Entry
 - OCR errors
 - Lack of complete information
 - Ambiguous data
 - Duplicate data
 - Data transformation errors
- What is the problem with data Quality:
 - Noise and outliers
 - Wrong data
 - Fake data
 - Missing values
 - Duplicate data

Agenda

- Data Quality Issues
 - Trouble with Data quality issues to Businesses and Industries
 - General Issues to all industries and businesses
 - Timeliness
 - Relevance
 - Knowledge about data
- Introduction to Data Processing

Data Quality Issues

- Have any business processes/activities been impacted by the data issue?
- If so, how many business processes/activities are impacted by the data issue?
- What business applications have failed as a result of the data issue?
- If so, how many business processes have failed?
- How many individuals are affected?
- How many systems are affected?
- What types of systems are affected?

Data Quality Issues

- • How many records are affected?
- • How many times has this issue been reported? Within what time frame?
- • How long has this been an issue?
- Then, based on the list of individuals and systems affected, the data quality analyst can review business impacts within the context of both known and newly discovered issues, asking questions such as these:

Data Quality issues

- • What are the potential business impacts?
- • Is this an issue that has already been anticipated based on the data requirements analysis process?
- • Has this issue introduced delays or halts in production information processing that must be performed within existing constraints?
- • Has this issue introduced delays in the development or deployment of critical business systems

Data Quality

- “Data of high quality if it suitable for intended use”-- Viewpoint
- Useful both Industries and business
- Statistical Field-Data Collect relevant to hypothesis.
- Many issues are there to specific application. Will discuss General Issues-
 - Timeliness
 - Relevance
 - Knowledge about data

Timeliness

- Consider the data from some ongoing phenomenon or process
- Snapshot of phenomenon
- Example-Purchasing Behavior of Customer / web browsing
- Snapshot → reality for Limited Time
- Data—Out of Date → Data Model /pattern is also out of date

Relevance

- The available data must contain the information necessary for the application.
- Example:
 - Predict Accident Rate of Driver-
 - Building Model
 - Information-age ,gender(Omitted then limited accuracy)
- Challenges- Sampling Bias
 - Sample does not contains different types of object in proportion to their actual occurrence in the population.
 - Example-
 - survey data-limited people involve- erroneous result-broader application

Knowledge about data

- Data set accompanied by Documentation
- Quality aid or hinder analysis
- Ex: documentation-redundant data information then strongly related
 - : Sale tax and purchase price
- Missing Value—9999 Analysis faulty
- Important characteristics of data quality
 - Precision of data
 - Data attributes
 - Scale of measurement
 - Origin of data

Data Transformation to make suitable for Mining

- Data Processing
 - Aggregation
 - Sampling
 - Dimensionality Reduction
- EDA Vs Classical Data Analysis

Summarized

- Data Quality Issue broader perspectives
- Data Quality General Issue-Timeliness,Relevance and Knowledge Data
- Introduction to Transformation of data

Lecture 6

EDA Vs Classical Data Analysis

Vaibhav Chunekar

20/1/2022

Agenda

- Recap Data Quality Issues
- Data Transformation Techniques
- EDA Vs Classical Data Analysis

Data Transformation in Data Mining

- The data are transformed in ways that are ideal for mining the data.
The data transformation involves steps that are:
 - **Smoothing**
 - **Aggregation**
 - **Discretization**
 - **Attribute Construction**
 - **Generalization**
 - **Normalization**

Smoothing

- It is a process that is used to remove noise from the dataset using some algorithms. It allows for highlighting important features present in the dataset. It helps in predicting the patterns. When collecting data, it can be manipulated to eliminate or reduce any variance or any other noise form.
- The concept behind data smoothing is that it will be able to identify simple changes to help predict different trends and patterns. This serves as a help to analysts or traders who need to look at a lot of data which can often be difficult to digest for finding patterns that they wouldn't see otherwise.

Aggregation

- Data collection or aggregation is the method of storing and presenting data in a summary format. The data may be obtained from multiple data sources to integrate these data sources into a data analysis description. This is a crucial step since the accuracy of data analysis insights is highly dependent on the quantity and quality of the data used. Gathering accurate data of high quality and a large enough quantity is necessary to produce relevant results.
- The collection of data is useful for everything from decisions concerning financing or business strategy of the product, pricing, operations, and marketing strategies.
- For **example**, Sales, data may be aggregated to compute monthly& annual total amounts.

Discretization

- It is a process of transforming continuous data into set of small intervals. Most Data Mining activities in the real world require continuous attributes. Yet many of the existing data mining frameworks are unable to handle these attributes.
- Also, even if a data mining task can manage a continuous attribute, it can significantly improve its efficiency by replacing a constant quality attribute with its discrete values.
- For **example**, (1-10, 11-20) (age:- young, middle age, senior).

Attribute Construction

- Where new attributes are created & applied to assist the mining process from the given set of attributes. This simplifies the original data & makes the mining more efficient

Generalization

- It converts low-level data attributes to high-level data attributes using concept hierarchy. For Example Age initially in Numerical form (22, 25) is converted into categorical value (young, old).
- For **example**, Categorical attributes, such as house addresses, may be generalized to higher-level definitions, such as town or country.

Normalization

- Data normalization involves converting all data variable into a given range.
- Techniques that are used for normalization are:
 - Min Max Normalization
 - Z Score Normalization
 - Decimal Scaling

Min-Max Normalization:

- This transforms the original data linearly.
- Suppose that min_A is the minima and max_A is the maxima of an attribute, P
- We Have the Formula:

$$v' = \frac{v - \text{min}_P}{\text{max}_P - \text{min}_P} (\text{new_max}_P - \text{new_min}_P) + \text{new_min}_P$$

- Where v is the value you want to plot in the new range.
- v' is the new value you get after normalizing the old value.

Solved example:

Suppose the minimum and maximum value for an attribute profit(P) are Rs. 10,000 and Rs. 100,000. We want to plot the profit in the range [0, 1]. Using min-max normalization the value of Rs. 20,000 for attribute profit can be plotted to:

$$v' = \frac{v - \text{min}_P}{\text{max}_P - \text{min}_P} (\text{new_max}_P - \text{new_min}_P) + \text{new_min}_P$$

$$\frac{20000 - 10000}{100000 - 10000} (1 - 0) + 0 = 0.11$$

And hence, we get the value of v' as 0.11

Z-Score Normalization:

- In z-score normalization (or zero-mean normalization) the values of an attribute (A), are normalized based on the mean of A and its standard deviation
- A value, v, of attribute A is normalized to v' by computing

$$v' = \frac{v - \bar{A}}{\sigma_A}$$

example:

- Let mean of an attribute P = 60, 000, Standard Deviation = 10, 000, for the attribute P. Using z-score normalization, a value of 85000 for P can be transformed to:

$$\frac{85000 - 60000}{10000} = 2.50$$

And hence we get the value of v' to be 2.5

Decimal Scaling:

- It normalizes the values of an attribute by changing the position of their decimal points
- The number of points by which the decimal point is moved can be determined by the absolute maximum value of attribute A.
- A value, v , of attribute A is normalized to v' by computing

- $$v' = \frac{v}{10^j}$$

where j is the smallest integer such that $\text{Max}(|v'|) < 1$

Example

- Suppose: Values of an attribute P varies from -99 to 99.
- The maximum absolute value of P is 99.
- For normalizing the values we divide the numbers by 100 (i.e., $j = 2$) or (number of integers in the largest number) so that values come out to be as 0.98, 0.97 and so on

What is EDA?

- Exploratory Data Analysis (EDA) is an approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to
 - maximize insight into a data set;
 - uncover underlying structure;
 - extract important variables;
 - detect outliers and anomalies;
 - test underlying assumptions;
 - develop parsimonious models; and
 - determine optimal factor settings.

Focus of EDA?

- The EDA approach is precisely that--an approach--not a set of techniques, but an attitude/philosophy about how a data analysis should be carried out.

Philosophy of EDA

- EDA is not identical to statistical graphics although the two terms are used almost interchangeably.
- Statistical graphics is a collection of techniques--all graphically based and all focusing on one data characterization aspect.
- EDA encompasses a larger venue; EDA is an approach to data analysis that postpones the usual assumptions about what kind of model the data follow with the more direct approach of allowing the data itself to reveal its underlying structure and model.
- EDA is not a mere collection of techniques; EDA is a philosophy as to how we dissect a data set; what we look for; how we look; and how we interpret.
- It is true that EDA heavily uses the collection of techniques that we call "statistical graphics", but it is not identical to statistical graphics per se.

History of EDA

- The seminal work in EDA is Exploratory Data Analysis, Tukey, (1977). Over the years it has benefitted from other noteworthy publications such as Data Analysis and Regression, Mosteller and Tukey (1977), Interactive Data Analysis, Hoaglin (1977), The ABC's of EDA, Velleman and Hoaglin (1981) and has gained a large following as "the" way to analyze a data set.

Techniques EDA

- Most EDA techniques are graphical in nature with a few quantitative techniques.
- The reason for the heavy reliance on graphics is that by its very nature the main role of EDA is to open-mindedly explore, and graphics gives the analysts unparalleled power to do so, enticing the data to reveal its structural secrets, and being always ready to gain some new, often unsuspected, insight into the data.
- In combination with the natural pattern-recognition capabilities that we all possess, graphics provides, of course, unparalleled power to carry this out.

Special Graphical Techniques in EDA

The particular graphical techniques employed in EDA are often quite simple, consisting of various techniques of:

1. Plotting the raw data (such as [data traces](#), [histograms](#), [bihistograms](#), [probability plots](#), [lag plots](#), [block plots](#), and [Youden plots](#)).
2. Plotting simple statistics such as [mean plots](#), [standard deviation plots](#), [box plots](#), and main effects plots of the raw data.
3. Positioning such plots so as to maximize our natural pattern-recognition abilities, such as using multiple plots per page.

EDA Vs classical Data Analysis

- These two approaches are similar in that they all start with a general science/engineering problem and all yield science/engineering conclusions. The difference is the sequence and focus of the intermediate steps.

classical Data Analysis	EDA
For classical analysis, the sequence is Problem => Data => Model => Analysis => Conclusions	For EDA, the sequence is Problem => Data => Analysis => Model => Conclusions

EDA Vs classical Data Analysis

For classical analysis, the data collection is followed by the imposition of a model (normality, linearity, etc.) and the analysis, estimation, and testing that follows are focused on the parameters of that model.

For EDA, the data collection is not followed by a model imposition; rather it is followed immediately by analysis with a goal of inferring what model would be appropriate.

Model

Classical Data Analysis	EDA
<p>The classical approach imposes models (both deterministic and probabilistic) on the data. Deterministic models include, for example, regression models and analysis of variance (ANOVA) models. The most common probabilistic model assumes that the errors about the deterministic model are normally distributed--this assumption affects the validity of the ANOVA F tests.</p>	<p>The Exploratory Data Analysis approach does not impose deterministic or probabilistic models on the data. On the contrary, the EDA approach allows the data to suggest admissible models that best fit the data.</p>

FOCUS

Classical Data Analysis	EDA
The two approaches differ substantially in focus. For classical analysis, the focus is on the model--estimating parameters of the model and generating predicted values from the model	For exploratory data analysis, the focus is on the data--its structure, outliers, and models suggested by the data

Techniques used

Classical Data Analysis	EDA
Classical techniques are generally <u>quantitative</u> in nature. They include <u>ANOVA</u> , <u>t tests</u> , <u>chi-squared tests</u> , and <u>F tests</u> .	EDA techniques are generally <u>graphical</u> . They include <u>scatter plots</u> , <u>character plots</u> , <u>box plots</u> , <u>histograms</u> , <u>bihistograms</u> , <u>probability plots</u> , <u>residual plots</u> , and <u>mean plots</u>

Rigor

Classical Data Analysis	EDA
<p>Classical techniques serve as the probabilistic foundation of science and engineering; the most important characteristic of classical techniques is that they are rigorous, formal, and "objective".</p>	<p>EDA techniques do not share in that rigor or formality. EDA techniques make up for that lack of rigor by being very suggestive, indicative, and insightful about what the appropriate model should be.</p> <p>EDA techniques are subjective and depend on interpretation which may differ from analyst to analyst, although experienced analysts commonly arrive at identical conclusions.</p>

Data Treatment

Classical Data Analysis	EDA
<p>Classical estimation techniques have the characteristic of taking all of the data and mapping the data into a few numbers ("estimates"). This is both a virtue and a vice. The virtue is that these few numbers focus on important characteristics (location, variation, etc.) of the population. The vice is that concentrating on these few characteristics can filter out other characteristics (skewness, tail length, autocorrelation, etc.) of the same population. In this sense there is a loss of information due to this "filtering" process.</p>	<p>The EDA approach, on the other hand, often makes use of (and shows) all of the available data. In this sense there is no corresponding loss of information.</p>

Data Treatment

Classical Data Analysis	EDA
<p>Classical estimation techniques have the characteristic of taking all of the data and mapping the data into a few numbers ("estimates"). This is both a virtue and a vice. The virtue is that these few numbers focus on important characteristics (location, variation, etc.) of the population. The vice is that concentrating on these few characteristics can filter out other characteristics (skewness, tail length, autocorrelation, etc.) of the same population. In this sense there is a loss of information due to this "filtering" process.</p> <p>.</p>	<p>The EDA approach, on the other hand, often makes use of (and shows) all of the available data. In this sense there is no corresponding loss of information.</p>

Assumptions

Classical Data Analysis	EDA
<p>Classical estimation techniques have the characteristic of taking all of the data and mapping the data into a few numbers ("estimates"). This is both a virtue and a vice. The virtue is that these few numbers focus on important characteristics (location, variation, etc.) of the population. The vice is that concentrating on these few characteristics can filter out other characteristics (skewness, tail length, autocorrelation, etc.) of the same population. In this sense there is a loss of information due to this "filtering" process.</p>	<p>Many EDA techniques make little or no assumptions--they present and show the data--all of the data--as is, with fewer encumbering assumptions</p>

Classical Data Analysis

1. Data collection is followed by the **imposition of a model** (normality, linearity) and the analysis, estimation, and testing that follows are focused on the parameters of that model.
2. Focus is on the **model**--estimating parameters of the model and generating predicted values from the model.
3. Problem => Data => Model => Analysis => Conclusions
4. Classical techniques are generally quantitative in nature. They include ANOVA, t tests, chi-squared tests, and F tests.
5. Classical estimation techniques have the characteristic of taking all of the data and mapping the data into a few numbers("estimates"). The virtue is that these few numbers focus on important characteristics(location, variation) of the population. The vice is that concentrating on these few characteristics can filter out other characteristics(skewness, autocorrelation, tail length) of the same population. In this sense there is a loss of information due to this "filtering" process.
6. In Data mining, the major areas of interest are clustering and anomaly detection and not exploratory.

EDA

1. Data collection is **not followed by a model imposition**; rather it is followed immediately by **analysis** with a goal of inferring what model would be appropriate.
2. Focus is on the **data**--its structure, outliers, and models suggested by the data.
3. Problem => Data => Analysis => Model => Conclusions
4. EDA techniques are generally graphical. They include scatter plots, character plots, box plots, histograms, probability plots, residual plots, and mean plots.
5. The EDA approach, on the other hand, often makes use of (and shows) all of the available data. In this sense there is no corresponding loss of information. Many EDA techniques make little or no assumptions--they present and show the data--all of the data--as is, with fewer encumbering assumptions.
6. In EDA the focus is on:
 - Summary statistics
 - Visualization
 - Online Analytical Processing (OLAP)

Classical Data Analysis

EDA



| Classical Data Analysis VS
EDA

EXAMPLE FOR EDA vs CLASSICAL DATA MINING

Consider a simple example of dataset which contains data of people belonging to Mumbai(Say 1 Million records).And your problem statement is to predict number of migrants flocking in Mumbai every year.

Classical analysis will include:

- 1.) Finding mean age of migrants
- 2.) From which state of India, maximum population is migrating in Mumbai
- 3.) Male to female ratio.
- 4.) Ordered list of for which domain, people are migrating?

EDA will include:

- 1.) Age-group wise migration of people.
- 2.) Histogram for Number of people based on Industry type
- 3.) Increase/decrease in Population of migrants year wise.

Summary

- Data Transformation Techniques
- EDA vs Classical Data Analysis

Lecture 7: Frequency Distribution and Measure of Central Tendency

Vaibhav Chunekar

24 Jan 2022

Frequency Distribution

- After collecting data, the first task for students used to organize and simplify the data so that it is possible to get a general overview of the results.
- This is the goal of descriptive statistical techniques.
- One method for simplifying and organizing data is to construct a **frequency distribution**.

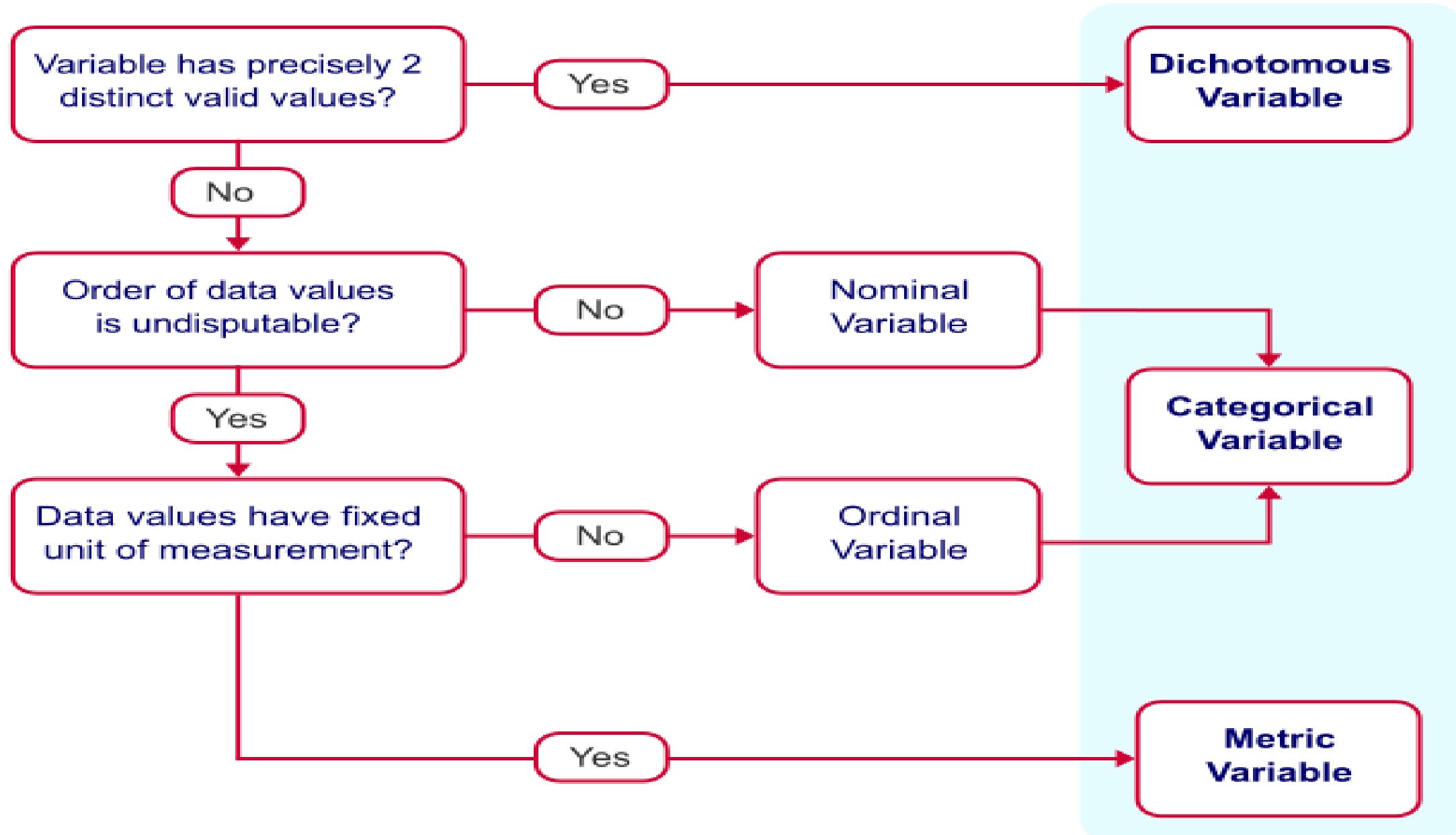
Frequency Distribution (cont.)

- A **frequency distribution** is an organized tabulation showing exactly how many individuals are located in each category on the scale of measurement.
- A frequency distribution presents an organized picture of the entire set of scores, and it shows where each individual is located relative to others in the distribution.

Frequency Distribution

- A **frequency distribution** is an overview of all distinct values in some variable and the number of times they occur.
- Frequency distribution tells “ How **frequencies** are **distributed** over values”.
- Frequency distributions are mostly used for summarizing **Categorical variables**.
- **Categorical variables** are variables on which calculations are not meaningful.
 - That's because metric variables tend to have many distinct values.
 - These result in huge tables and charts that don't give insight into your data.
 - In this case,histogram are the way to go as they visualize frequencies for *intervals* of values rather than each distinct value.

MEASUREMENT LEVELS - MODERN APPROACH



Frequency Distribution Tables

- A **frequency distribution table** consists of at least two columns - one listing categories on the scale of measurement (X) and another for frequency (f).
- In the X column, values are listed from the highest to lowest, without skipping any.
- For the frequency column, tallies are determined for each value (how often each X value occurs in the data set). These tallies are the frequencies for each X value.
- The sum of the frequencies should equal N .

Frequency Distribution Tables (cont.)

- A third column can be used for the proportion (p) for each category: $p = f/N$. The sum of the p column should equal 1.00.
- A fourth column can display the percentage of the distribution corresponding to each X value. The percentage is found by multiplying p by 100. The sum of the percentage column is 100%.

Regular Frequency Distribution

- When a frequency distribution table lists all of the individual categories (X values) it is called a **regular frequency distribution**.

Frequency Distribution - Example

	id	fname	sex	major
1	7042	Piper	female	Sociology
2	7104	Nicole	female	Anthropology
3	8016	Samuel	male	Other
4	8088	Logan	male	Psychology
5	8100	Alexa	female	Anthropology
6	9002	Scarlett	female	Sociology
7	9035	Wyatt	male	Economy
8

Frequency Distribution -183 Student (as per subject interest)

What's currently your (primary) major?	N	Percent
Psychology	62	33.9%
Economy	35	19.1%
Sociology	33	18.0%
Anthropology	37	20.2%
Other	16	8.7%
Total	183	100.0%

FREQUENCIES →
ARE DISTRIBUTED OVER
← VALUES

Relative Frequencies

- Frequencies *relative to* (divided by) the total number of values.
Relative frequencies are often shown as percentages or proportions.

What's currently your (primary) major?	N	Percent
Psychology	62	33.9%
Economy	35	19.1%
Sociology		18.0%
Anthropology		20.2%
Other	16	8.7%
Total	183	100.0%

RELATIVE
FREQUENCIES ←

Real Sample _Computer Hardware data_1987

Sr.No.	vendor name	Model Name	machine cycle time	minimum main memory	maximum main memory	cache memory	minimum channels	maximum channels	published relative performance	estimated relative performance
1	adviser	32/60	125	256	6000	256	16	128	198	199
2	amdahl	470v/7	29	8000	32000	32	8	32	269	253
3	amdahl	470v/7a	29	8000	32000	32	8	32	220	253
4	amdahl	470v/7b	29	8000	32000	32	8	32	172	253
5	amdahl	470v/7c	29	8000	16000	32	8	16	132	132

Computer Hardware Vendors-Data Statistics

adviser, amdahl, apollo, basf, bti, burroughs, c.r.d, cambex, cdc, dec, dg, formation, four-phase, gould, honeywell, hp, ibm, ipl, magnuson, microdata, nas, ncr, nixdorf, perkin-elmer, prime, siemens, sperry, sratus, wang : (Total 30 vendors)(Total Sample 209)

Frequency Distribution- 209 instances

Sr.No	Vendor Name	Total Machine
1	adviser	1
2	amdahl	9
3	apollo	2
4	basf	2
5	bti	2
6	burroughs	8
7	c.r.d	6
8	cdc	9
9	cambex	5
10	dec	6
11	dg	7
12	formation	5
13	four-phase	1
14	Gould	3
15	hp	7

Sr.No	Vendor Name	Total Machine
1	harris	7
2	honeywell	13
3	ibm	32
4	ipl	6
5	magnuson	6
6	microdata	1
7	nas	19
8	ncr	13
9	nixdorf	3
10	perkin-elmer	3
11	prime	5
12	siemens	12
13	sperry	13
14	sstatus	1
15	wang	2

Relative Frequency

Vendor Name	Frequency	Relative Frequency(209)
ibm	32	15.31100478
nas	19	9.090909091
honeywell	13	6.220095694
ncr	13	6.220095694
sperry	13	6.220095694
siemens	12	5.741626794
amdahl	9	4.306220096
cdc	9	4.306220096
burroughs	8	3.827751196

bASF	2	0.956937799
BTI	2	0.956937799
WANG	2	0.956937799
adviser	1	0.4784689
four-phase	1	0.4784689
microdata	1	0.4784689
sstatus	1	0.4784689
	209	100%

Cumulative Frequency

- *Cumulative relative frequency* is the accumulation of the previous relative frequencies. To find the cumulative relative frequencies, add all the previous relative frequencies to the relative frequency for the current row. \top
- Twenty students were asked how many hours they worked per day. Their responses, in hours, are as follows:
- 5; 6; 3; 3; 2; 4; 7; 5; 2; 3; 5; 6; 5; 4; 4; 3; 5; 2; 5; 3.

Frequency Distributions

DATA VALUE	FREQUENCY
2	3
3	5
4	3
5	6
6	2
7	1

Relative Frequency: Total student 20

DATA VALUE	FREQUENCY	Rel Freq
2	3	$3/20*100=15$
3	5	$5/20*100=25$
4	3	15
5	6	30
6	2	10
7	1	5
	20	100%

Frequency Table of Student Work Hours with Relative and Cumulative Relative Frequencies

DATA VALUE	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
2	3	0.15	0-0.15
3	5	0.25	0.15 + 0.25 = 0.40
4	3	0.15	0.40 + 0.15 = 0.55
5	6	0.30	0.55 + 0.30 = 0.85
6	2	0.10	0.85 + 0.10 = 0.95
7	1	0.05	0.95 + 0.05 = 1.00

Grouped Frequency Distribution

- A grouped frequency distribution is a table to organize data in which the data are grouped into classes with more than one unit in width. Used when the data is large, or it makes sense to group the data.

Grouped Frequency Distribution

- Sometimes, however, a set of scores covers a wide range of values. In these situations, a list of all the X values would be quite long - too long to be a “simple” presentation of the data.
- To remedy this situation, a **grouped frequency distribution** table is used.

Grouped Frequency Distribution (cont.)

- In a grouped table, the X column lists groups of scores, called **class intervals**, rather than individual values.
- These intervals all have the same width, usually a simple number such as 2, 5, 10, and so on.
- Each interval begins with a value that is a multiple of the interval width. The interval width is selected so that the table will have approximately ten intervals.

Frequency Table of Soccer Player Height

Heights (Inches)	Frequency	Relative Frequency	Cumulative Relative Frequency
59.95–61.95	5	$5/100=0.05$	0.05
61.95–63.95	3	$3/100=0.03$	$0.05+0.03=0.08$
63.95–65.95	15	$15/100=0.15$	$0.08+0.15=0.23$
65.95–67.95	40	$40/100=0.40$	$0.23+0.40=0.63$
67.95–69.95	17	$17/100=0.17$	$0.63+0.17=0.80$
69.95–71.95	12	$12/100=0.12$	$0.80+0.12=0.92$
71.95–73.95	7	$7/100=0.07$	$0.92+0.07=0.99$
73.95–75.95	1	$1/100=0.01$	$0.99+0.01=1.00$
	Total = 100	Total = 1.00	

Group Frequency Distributions

In this sample, there are

five players whose heights fall within the interval 59.95–61.95 inches,

three players whose heights fall within the interval 61.95–63.95 inches,

15 players whose heights fall within the interval 63.95–65.95 inches,

40 players whose heights fall within the interval 65.95–67.95 inches,

17 players whose heights fall within the interval 67.95–69.95 inches,

12 players whose heights fall within the interval 69.95–71.95,

seven players whose heights fall within the interval 71.95–73.95, and

one player whose heights fall within the interval 73.95–75.95.

All heights fall between the endpoints of an interval and not at the endpoints.

Measures of central tendency

Measure of central tendency

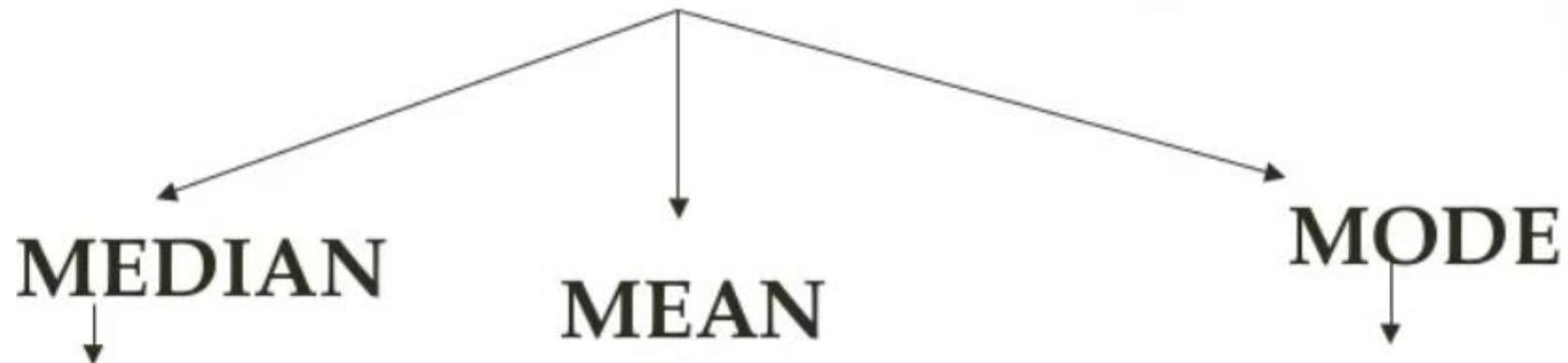
- A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data.
- As such, measures of central tendency are sometimes called measures of central location.
- They are also classed as summary statistics.
- The mean (often called the average) is most likely the measure of central tendency that you are most familiar with, but there are others, such as the median and the mode.
- The mean, median and mode are all valid measures of central tendency, but under different conditions, some measures of central tendency become more appropriate to use than others. In the following sections, we will look at the mean, mode and median, and learn how to calculate them and under what conditions they are most appropriate to be used.

Measures of central tendency

The *mean*, *median*, and *mode* are measures of central tendency—values that describe the center of a data set.

- The *mean* is the sum of the values in the set divided by the number of values. It is often represented as x .
- The *median* is the middle value or the mean of the two middle values when the set is ordered numerically. The
- *mode* is the value or values that occur most often. A data set may have one mode, no mode, or several modes.

MEASURE OF CENTRAL TENDANCY



MEDIAN
The middle value of the
data

MEAN

The average of the data

MODE

most commonly
occurring value

Mean (Arithmetic)

- The mean (or average) is the most popular and well known measure of central tendency.
- It can be used with both discrete and continuous data.
- The mean is equal to the sum of all the values in the data set divided by the number of values in the data set. So, if we have n values in a data set and they have values

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

$$\bar{x} = \frac{\sum x}{n}$$

$$\mu = \frac{\sum x}{n}$$

Pros of Mean

- The mean is essentially a model of your data set. It is the value that is most common.
- You will notice, however, that the mean is not often one of the actual values that you have observed in your data set.
- However, one of its important properties is that it minimises error in the prediction of any one value in your data set.
- That is, it is the value that produces the lowest amount of error from all other values in the data set.
- An important property of the mean is that it includes every value in your data set as part of the calculation.
- In addition, the mean is the only measure of central tendency where the sum of the deviations of each value from the mean is always zero

Cons of mean

1. It is particularly susceptible to the influence of outliers. These are values that are unusual compared to the rest of the data set by being especially small or large in numerical value.
2. Take Example wages of staff at a factory below:

Staff	1	2	3	4	5	6	7	8	9	10
Salary	15k	18k	16k	14k	15k	15k	12k	17k	90k	95k

3. The mean salary for these ten staff is \$30.7k. However, inspecting the raw data suggests that this mean value might not be the best way to accurately reflect the typical salary of a worker, as most workers have salaries in the \$12k to 18k range.
4. The mean is being skewed by the two large salaries. Therefore, in this situation, we would like to have a better measure of central tendency. As we will find out later, taking the median would be a better measure of central tendency in this situation

Median

- The median is the middle score for a set of data that has been arranged in order of magnitude.
- The median is less affected by outliers and skewed data.

Outliers

- An **outlier** is an extreme value that is much less than or much greater than the other data values. Outliers have a strong effect on the mean and standard deviation.
- If an outlier is the result of measurement error or represents data from the wrong population, it is usually removed. There are different ways to
- determine whether a value is an outlier. One is to look for data values that are more than 3 standard deviations from the mean

Median

- The median is the middle score for a set of data that has been arranged in order of magnitude.
- The median is less affected by outliers and skewed data.
- Compute Median of following data:

65	55	89	56	35	14	56	55	87	45	92
----	----	----	----	----	----	----	----	----	----	----

.

We first need to rearrange that data into order of magnitude (smallest first):

14	35	45	55	55	56	56	65	87	89	92
----	----	----	----	----	-----------	----	----	----	----	----

Median Computation continue..

65	55	89	56	35	14	56	55	87	45
----	----	----	----	----	----	----	----	----	----

We again rearrange that data into order of magnitude (smallest first):

14	35	45	55	55	56	56	65	87	89
----	----	----	----	----	----	----	----	----	----

Now take the 5th and 6th score in our data set and average them to get a median of 55.5

Mode

- Mode means a value or a number that appears most frequently in a dataset.
- **Unimodal List:** A list of given data with only one mode is called a unimodal list.
- **Bimodal List:** A list of given data with two modes is called a bimodal list

Mode Formula

$$\text{Mode} = L + h \frac{(f_m - f_1)}{(f_m - f_1) + (f_m - f_2)}$$

- L is the lower limit of the modal class
- h is the size of the class interval
- Fm is the frequency of the modal class
- f1 is the frequency of the class preceding the modal class
- f2 is the frequency of the class succeeding the modal class

Example 1: Finding Measures of Central Tendency

**Find the mean, median, and mode of the data.
deer at a feeder each hour: 3, 0, 2, 0, 1, 2, 4**

Mean: $\frac{3+0+2+0+1+2+4}{7} = \frac{12}{7} \approx 1.7 \text{ deer}$

Median: 0 0 1 **2** 2 3 4 = 2 deer

Mode: The most common results are 0 and 2.

Find the mean, median, and mode of the data set.

{2, 5, 6, 2, 6}

Mean: $\frac{2+5+6+2+6}{5} = \frac{21}{5} = 4.2$

Median: 2 2 **5** 6 6 = 5

Mode: 2 and 6

A *weighted average* is a mean calculated by using frequencies of data values. Suppose that 30 movies are rated as follows:

Movie Ratings					
Rating	★★★★	★★★	★★	★	no stars
Number of Movies	8	12	7	2	1

weighted average of stars =

$$\frac{8(4) + 12(3) + 7(2) + 2(1) + 1(0)}{8 + 12 + 7 + 2 + 1} = \frac{84}{30} = 2.8 \text{ stars}$$

For numerical data, the weighted average of all of those outcomes is called the **expected value** for that experiment.

The **probability distribution** for an experiment is the function that pairs each outcome with its probability.

Example 2: Finding Expected Value

The probability distribution of successful free throws for a practice set is given below. Find the expected number of successes for one set.

Number of Good Free Throws, n	0	1	2	3
Prob. of n Good Free Throws	$\frac{3}{20}$	$\frac{3}{20}$	$\frac{1}{5}$	$\frac{1}{2}$

Example 2 Continued

$$\text{expected value} = 0\left(\frac{3}{20}\right) + 1\left(\frac{3}{20}\right) + 2\left(\frac{1}{5}\right) + 3\left(\frac{1}{2}\right)$$

Use the weighted average.

$$= 0 + \frac{3}{20} + \frac{2}{5} + \frac{3}{2}$$

Simplify.

$$= 0 + \frac{3}{20} + \frac{8}{20} + \frac{30}{20} = \frac{41}{20} = 2.05$$

The expected number of successful free throws is 2.05.

Another Example

The probability distribution of the number of accidents in a week at an intersection, based on past data, is given below. Find the expected number of accidents for one week.

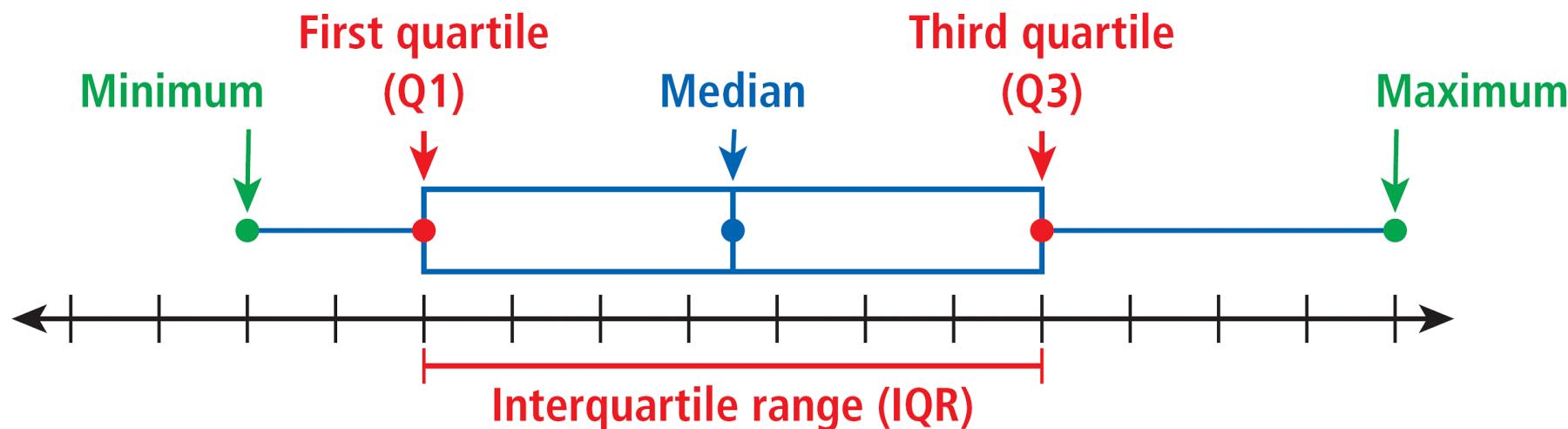
Number of accidents n	0	1	2	3
Probability of n accidents	0.75	0.15	0.08	0.02

Use the weighted average.

$$\begin{aligned}\text{expected value} &= 0(0.75) + 1(0.15) + 2(0.08) + 3(0.02) \\ &= 0.37 \quad \text{\textcolor{blue}{Simplify}.}\end{aligned}$$

The expected number of accidents is 0.37.

A *box-and-whisker plot* shows the spread of a data set. It displays 5 key points: the **minimum** and **maximum** values, the **median**, and the **first** and **third quartiles**.



The quartiles are the medians of the lower and upper halves of the data set. If there are an odd number of data values, do not include the median in either half.

The *interquartile range*, or IQR, is the difference between the 1st and 3rd quartiles, or $Q3 - Q1$. It represents the middle 50% of the data.

Example 3: Making a Box-and-Whisker Plot and Finding the Interquartile Range

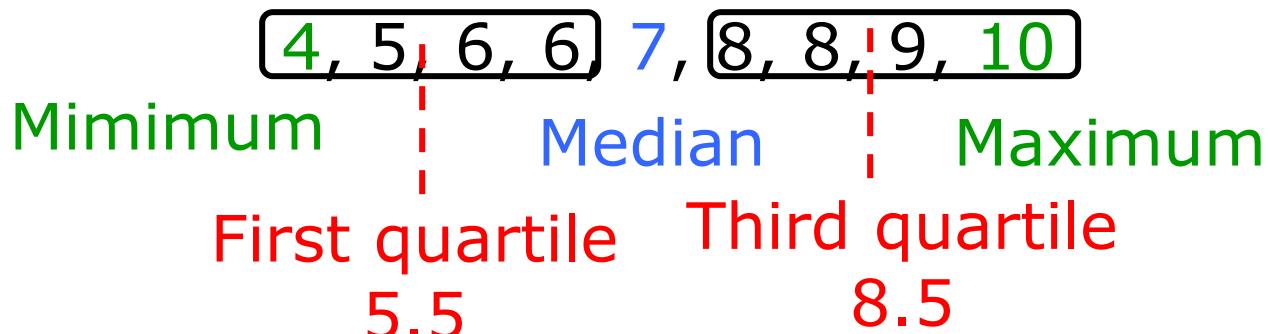
Make a box-and-whisker plot of the data. Find the interquartile range.

{6, 8, 7, 5, 10, 6, 9, 8, 4}

Step 1 Order the data from least to greatest.

4, 5, 6, 6, 7, 8, 8, 9, 10

Step 2 Find the minimum, maximum, median, and quartiles.

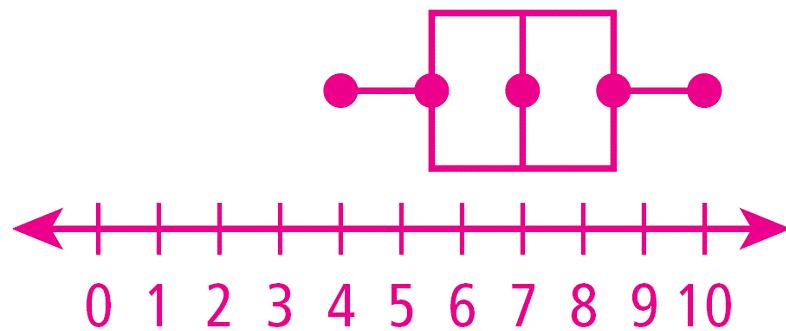


Example 3 Continued

Step 3 Draw a box-and-whisker plot.

Draw a number line, and plot a point above each of the five values. Then draw a box from the first quartile to the third quartile with a line segment through the median. Draw whiskers from the box to the minimum and maximum.

Example 3 Continued



$$\text{IRQ} = 8.5 - 5.5 = 3$$

The interquartile range is 3, the length of the box in the diagram.

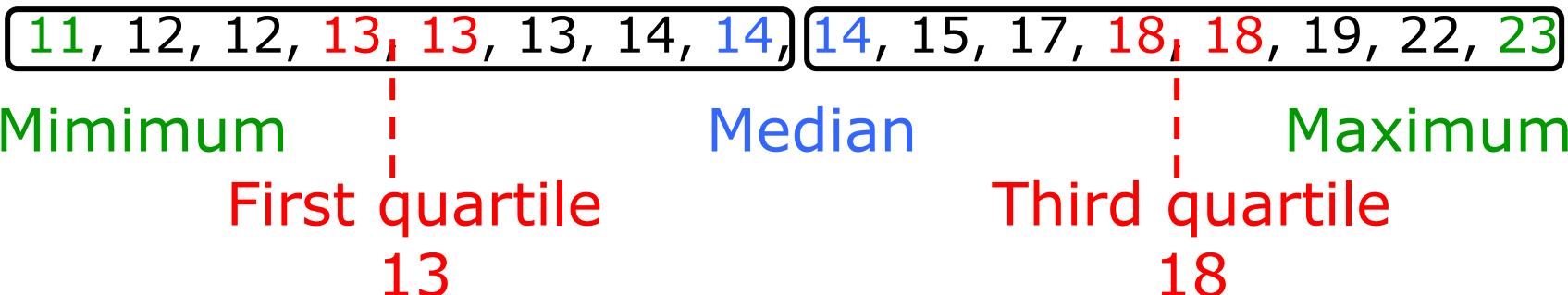
Example 3

Make a box-and-whisker plot of the data. Find the interquartile range. {13, 14, 18, 13, 12, 17, 15, 12, 13, 19, 11, 14, 14, 18, 22, 23}

Step 1 Order the data from least to greatest.

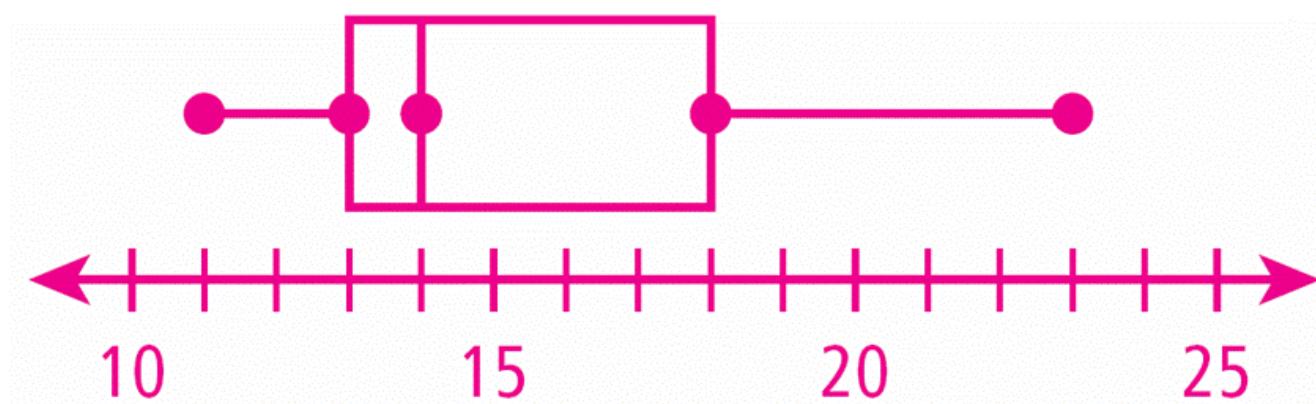
11, 12, 12, 13, 13, 13, 14, 14, 14, 14, 15, 17, 18, 18, 19, 22, 23

Step 2 Find the minimum, maximum, median, and quartiles.



Example 3 Continued

Step 3 Draw a box-and-whisker plot.



$$\text{IQR} = 18 - 13 = 5$$

The interquartile range is 5, the length of the box in the diagram.

The data sets **{19, 20, 21}** and **{0, 20, 40}** have the same mean and median, but the sets are very different. The way that data are spread out from the mean or median is important in the study of statistics.

A *measure of variation* is a value that describes the spread of a data set. The most commonly used measures of variation are the *range*, the interquartile range, the *variance*, and the *standard deviation*.

The **variance**, denoted by σ^2 , is the average of the squared differences from the mean. **Standard deviation**, denoted by σ , is the square root of the variance and is one of the most common and useful measures of variation.

Low standard deviations indicate data that are clustered near the measures of central tendency, whereas high standard deviations indicate data that are spread out from the center.

Finding Variance and Standard Deviation

Step 1. Find the mean of the data, \bar{x} .

Step 2. Find the difference between the mean and each data value, and square it.

Step 3. Find the variance, σ^2 , by adding the squares of all of the differences from the mean and dividing by the number of data values.

Step 4. Find the standard deviation, σ , by taking the square root of the variance.

Example 4: Finding the Mean and Standard Deviation

Find the mean and standard deviation for the data set of the number of people getting on and off a bus for several stops.

{6, 8, 7, 5, 10, 6, 9, 8, 4}

Step 1 Find the mean.

$$\bar{x} = \frac{6+8+7+5+10+6+9+8+4}{9} = 7$$

Example 4 Continued

Step 2 Find the difference between the mean and each data value, and square it.

Data value x	6	8	7	5	10	6	9	8	4
$x - \bar{x}$	-1	1	0	-2	3	-1	2	1	-3
$(x - \bar{x})^2$	1	1	0	4	9	1	4	1	9

Find the mean and standard deviation for the data set of the number of elevator stops for several rides.

{0, 3, 1, 1, 0, 5, 1, 0, 3, 0}

Step 1 Find the mean.

$$\bar{x} = \frac{0+3+1+1+0+5+1+0+3+0}{10} = 1.4$$

Step 2 Find the difference between the mean and each data value, and square it.

Data Value x	0	3	1	1	0	5	1	0	3	0
$x - \bar{x}$	-1.4	1.6	-0.4	-0.4	-1.4	3.6	-0.4	-1.4	1.6	-1.4
$(x - \bar{x})^2$	1.96	2.56	0.16	0.16	1.96	12.96	0.16	1.96	2.56	1.96

Step 3 Find the variance.

Find the average of the last row of the table

$$\sigma^2 = \frac{1.96 + 2.56 + 0.16 + 0.16 + 1.96 + 12.96 + 0.16 + 1.96 + 2.56 + 1.96}{10}$$

$$\sigma^2 = 2.64$$

Step 4 Find the standard deviation.

$$\sigma = \sqrt{2.64} \approx 1.6$$

The standard deviation is the square root of the variance.

The mean is 1.4 stops and the standard deviation is about 1.6 stops.

Lecture -8

Prediction of Missing Value with Covariance and Correlation

Vaibhav Chunekar

27 January 2022

Covariance

- In statistics and probability theory, covariance deals with the joint variability of two random variables: x and y.
- Generally, it is treated as a statistical tool used to define the relationship between two variables.
- **Covariance** is a measure of the relationship between two random variables and to what extent, they change together. Or we can say, in other words, it defines the changes between the two variables, such that change in one variable is equal to change in another variable.
- This is the property of a function of maintaining its form when the variables are linearly transformed.
- Covariance is measured in units, which are calculated by multiplying the units of the two variables

Types of Covariance

- **Positive Covariance**
- If the covariance for any two variables is positive, that means, both the variables move in the same direction. Here, the variables show similar behaviour. That means, if the values (greater or lesser) of one variable corresponds to the values of another variable, then they are said to be in positive covariance.
- **Negative Covariance**
- If the covariance for any two variables is negative, that means, both the variables move in the opposite direction. It is the opposite case of positive covariance, where greater values of one variable correspond to lesser values of another variable and vice-versa

Covariance formula

- Covariance formula is a statistical formula, used to evaluate the relationship between two variables.
- It is one of the statistical measurements to know the relationship between the variance between the two variables.
- Let us say X and Y are any two variables, whose relationship has to be calculated.
- Thus the covariance of these two variables is denoted by $\text{Cov}(X,Y)$.

Covariance Formula

Population Covariance Formula

$$\mathbf{Cov(x,y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N}}$$

Sample Covariance

$$\mathbf{Cov(x,y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N - 1}}$$

Where,

- x_i = data value of x
- y_i = data value of y
- \bar{x} = mean of x
- \bar{y} = mean of y
- N = number of data values.

Example: Calculate the coefficient of covariance for the following data:

X	2	8	18	20	28	30
Y	5	12	18	23	45	50

Number of observations = 6

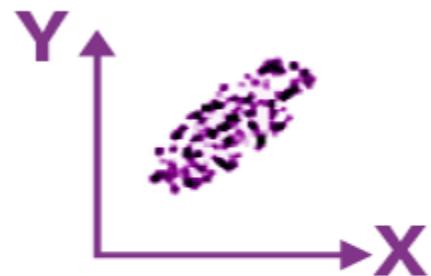
Mean of X = 17.67

Mean of Y = 25.5

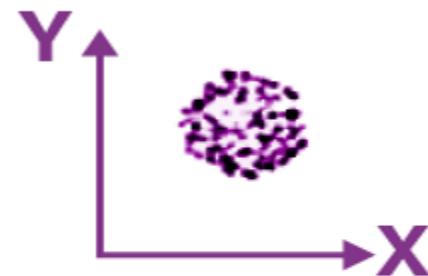
$$\text{Cov}(X, Y) = \frac{1}{6} [(2 - 17.67)(5 - 25.5) + (8 - 17.67)(12 - 25.5) + (18 - 17.67)(18 - 25.5) + (20 - 17.67)(23 - 25.5) + (28 - 17.67)(45 - 25.5) + (30 - 17.67)(50 - 25.5)]$$

$$= 157.83$$

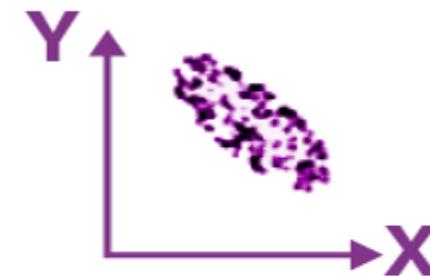
Interpretations Covariance



$$\text{cov}(X, Y) > 0$$



$$\text{cov}(X, Y) \approx 0$$



$$\text{cov}(X, Y) < 0$$

Interpretations Covariance

- If $\text{cov}(X, Y)$ is greater than zero, then we can say that the covariance for any two variables is positive and both the variables move in the same direction.
- If $\text{cov}(X, Y)$ is less than zero, then we can say that the covariance for any two variables is negative and both the variables move in the opposite direction.
- If $\text{cov}(X, Y)$ is zero, then we can say that there is no relation between two variables.

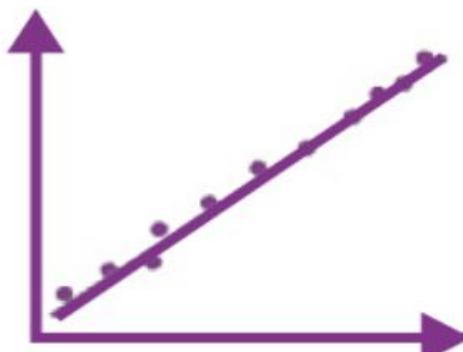
Correlation Coefficient & Formula

- Correlation estimates the depth of the relationship between variables. It is the estimated measure of covariance and is dimensionless. In other words, the correlation coefficient is a constant value always and does not have any units. The relationship between the correlation coefficient and covariance is given by:

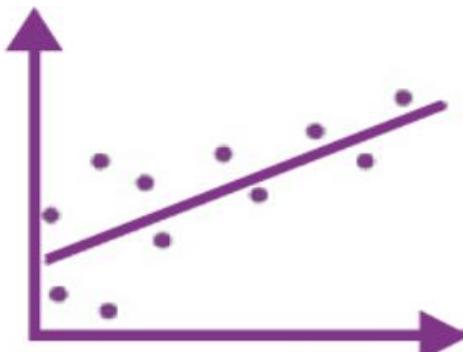
$$\text{Correlation}, \rho(X,Y) = \text{Cov}(X,Y)/\sigma_X \sigma_y$$

-
- $\rho(X,Y)$ = correlation between the variables X and Y
 - $\text{Cov}(X,Y)$ = covariance between the variables X and Y
 - σ_X = standard deviation of the X variable
 - σ_Y = standard deviation of the Y variable

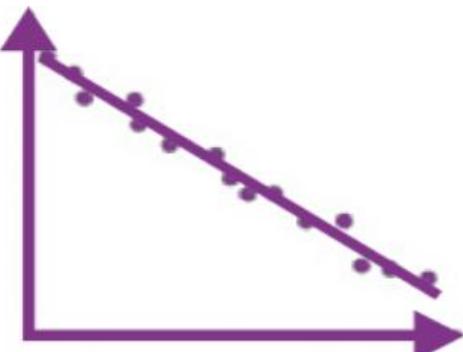
Graphical representation of correlation among two variables



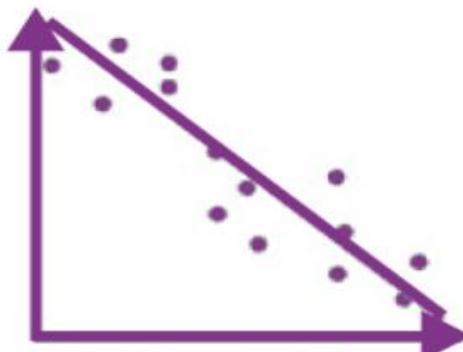
Strong positive correlation



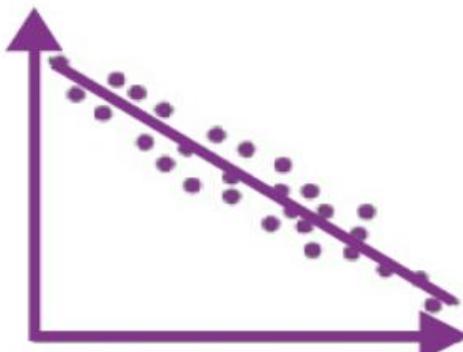
Weak positive correlation



Strong negative correlation



Weak negative correlation



Moderate negative correlation



No correlation

Difference Between Covariance-Correlation

Covariance	Correlation
<p>It is a measure to show the extent to which given two random variables change with respect to each other.</p>	<p>It is a measure used to describe how strongly the given two random variables are related to each other.</p>
<p>It is a measure of correlation.</p>	<p>It is defined as the scaled form of covariance.</p>
<p>The value of covariance lies between $-\infty$ and $+\infty$.</p>	<p>The value of correlation lies between -1 and +1.</p>
<p>It indicates the direction of the linear relationship between the given two variables.</p>	<p>It measures the direction and strength of the linear relationship between the given two variables.</p>

Data Analysis Process

- Everyone doing analysis has some missing data, especially survey researchers, market researchers, database analysts, researchers and social scientists.
- Missing data are questions without answers or variables without observations.
- Even a small percent of missing data can cause serious problems with your analysis leading you to draw wrong conclusions. Real-world databases are highly susceptible to noise, missing, and inconsistent data due to they are typically huge in size often in gigabytes or more.
- We have to preprocess the data in order to help improve to quality of data and so as to improve the efficiency and ease of mining access. There are number of data preprocessing techniques.
- Data cleaning can be applied to remove noise and correct inconsistencies in the data. Data integration merges data from multiple sources into a coherent data store, such as a data warehouse or a data cube.
- Data transformations, such as normalization, may be applied. Data reduction can reduce the data size by aggregating, eliminating redundant features, or clustering, for instance

Need of preprocessing data:

- The data you wish to analyze by data mining techniques are incomplete (lacking attribute values or certain attributes of interest), noisy (containing errors) and inconsistent.
- Incomplete data can occur in many reasons.
- Attribute values may not be available, not considering important at the time of entry.
- Missing data, particularly tuples with missing values for some attributes, may need to be inferred

Data cleaning:

- Real world data tend to be noisy, incomplete, and inconsistent. Data cleaning routines attempt to fill in missing values, smooth out noise while identifying outliers and correct inconsistencies in the data.
- We concentrate mainly on filling of missing values by ignoring the data row completely, filling the missing values manually, use the global constant to fill the missing values, use the attribute mean for 1 column of data, same using to fill all columns of data, using most probable value to fill missing value (Regression algorithm).
- In the regression method, a regression model is fitted for each variable with missing values. Based on the resulting model, a new regression model is then drawn and is used to impute the missing values for the variable. Since the data set has a several missing data patterns, the process is repeated sequentially for variables with missing values.

How to deal Missing Values?

- We have to fill those missing data cells with 6 possible ways.
- 1. Ignoring the data row completely
- 2. Filling missing values manually
- 3. Use a global constant to fill the missing values
- 4. Use the attribute mean to fill the missing value
- 5. Use the attribute mean for all samples belonging to the same class as the given tuple
- 6. Use the most probable value to fill the missing value (Predicting by Regression algorithm)

Regression Methodology

- A regression is a statistical analysis, assessing the association between two variables.
- It is used to find the relationship between two variables.
- Regression Formula:
- Regression Equation (y) = $a + bx$
 - slope -'b',
 - Intercept- 'a'
 - r-coefficient of correlation
 - x and y are the variables.
 - b = the slope of the regression line
 - a = the intercept point of the regression line and the y axis.
 - N = Number of values or elements
 - X = First Score
 - Y = Second Score
 - ΣXY = Sum of the product of first and Second Scores
 - ΣX = Sum of First Scores ΣY = Sum of Second Scores
 - ΣX^2 = Sum of square First Scores

$$\text{Intercept}(a) = \frac{\Sigma Y - b(\Sigma X)}{N}$$

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

Formula

$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$$

$$b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

How to find Linear Regression ?

X Values	Y Values
60	3.1
61	3.6
62	3.8
63	4
65	4.1

First we will find slope,
intercept and
use it to form regression equation

Step by Step approach:

Step 1. Count the number of values. $n = 5$

Step 2: Find XY , X^2

X Value	Y Value	X*Y	X*X
60	3.1	$60 * 3.1 = 186$	$60 * 60 = 3600$
61	3.6	$61 * 3.6 = 219.6$	$61 * 61 = 3721$
62	3.8	$62 * 3.8 = 235.6$	$62 * 62 = 3844$
63	4	$63 * 4 = 252$	$63 * 63 = 3969$
65	4.1	$65 * 4.1 = 266.5$	$65 * 65 = 4225$

- Step 3: Find ΣX , ΣY , ΣXY , ΣX^2 .
 - $\Sigma X = 311$
 - $\Sigma Y = 18.6$
 - $\Sigma XY = 1159.7$
 - $\Sigma X^2 = 19359$
- Step 4: Substitute the above information in slope formula(Compute b):

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

- $((5)*(1159.7) - (311)*(18.6)) / ((5)*(19359) - (311)^2)$
- $= (5798.5 - 5784.6) / (96795 - 96721)$
- $= 13.9 / 74$
- **b= 0.19**

- Step 5: Compute Intercept or a by fitting the value in the formula:

- $\Sigma X = 311$
- $\Sigma Y = 18.6$
- $\Sigma XY = 1159.7$
- $\Sigma X^2 = 19359$

$$\text{Intercept}(a) = \frac{\Sigma Y - b(\Sigma X)}{N}$$

- Then
- $= (18.6 - 0.19(311))/5$
- $= (18.6 - 59.09)/5$
- $= -40.49/5$
- $= -8.098$

- Step 6: Then substitute these values in regression equation formula
- Regression Equation(y) = $a+b(x)$
 - $= -8.098 + 0.19x.$
- Example of Prediction with Regression:
- Suppose if we want to know the approximate y value for the variable $x = 64$.
- Then we can substitute the value in the above equation.
- Regression Equation(y) = $a + bx = -8.098 + 0.19(64).$
 - $= -8.098 + 12.16$
 - $= 4.06$

Prediction of Value with Linear Regression

X Value	Y Value
60	3.1
61	3.6
62	3.8
63	4
65	4.1
64	4.06

Multiple linear regression

- **Multiple linear regression** is a method we can use to quantify the relationship between two or more predictor variables and a **response variable**.

Multiple Linear Regression using Paper Pen

- Suppose we have the following dataset with one response variable y and two predictor variables X_1 and X_2

y	x_1	x_2
140	60	22
155	62	25
159	67	24
179	70	20
192	71	15
200	72	14
212	75	14
215	78	11

- Step 1: Calculate X_1^2 , X_2^2 , X_1y , X_2y and X_1X_2

y	X_1	X_2
140	60	22
155	62	25
159	67	24
179	70	20
192	71	15
200	72	14
212	75	14
215	78	11
Mean	181.5	69.375
Sum	1452	555
		18.125
		145

X_1^2	X_2^2	X_1y	X_2y	X_1X_2
3600	484	8400	3080	1320
3844	625	9610	3875	1550
4489	576	10653	3816	1608
4900	400	12530	3580	1400
5041	225	13632	2880	1065
5184	196	14400	2800	1008
5625	196	15900	2968	1050
6084	121	16770	2365	858
38767	2823	101895	25364	9859

Sum

- **Step 2: Calculate Regression Sums.**

- $\sum X_1^2 = \sum X_1^2 - (\sum X_1)^2 / n = 38,767 - (555)^2 / 8 = \mathbf{263.875}$
- $\sum X_2^2 = \sum X_2^2 - (\sum X_2)^2 / n = 2,823 - (145)^2 / 8 = \mathbf{194.875}$
- $\sum X_1y = \sum X_1y - (\sum X_1 \sum y) / n = 101,895 - (555 * 1,452) / 8 = \mathbf{1,162.5}$
- $\sum X_2y = \sum X_2y - (\sum X_2 \sum y) / n = 25,364 - (145 * 1,452) / 8 = \mathbf{-953.5}$
- $\sum X_1X_2 = \sum X_1X_2 - (\sum X_1 \sum X_2) / n = 9,859 - (555 * 145) / 8 = \mathbf{-200.375}$

y	X₁	X₂
140	60	22
155	62	25
159	67	24
179	70	20
192	71	15
200	72	14
212	75	14
215	78	11
Mean	181.5	69.375
Sum	1452	555
		145

	X₁²	X₂²	X₁y	X₂y	X₁X₂
	3600	484	8400	3080	1320
	3844	625	9610	3875	1550
	4489	576	10653	3816	1608
	4900	400	12530	3580	1400
	5041	225	13632	2880	1065
	5184	196	14400	2800	1008
	5625	196	15900	2968	1050
	6084	121	16770	2365	858
Sum	38767	2823	101895	25364	9859

Reg Sums	263.875	194.875	1162.5	-953.5	-200.375

- Step 3: Calculate b_0 , b_1 , and b_2
- The formula to calculate b_1 is:
$$[(\sum x_2^2)(\sum x_1y) - (\sum x_1x_2)(\sum x_2y)] / [(\sum x_1^2)(\sum x_2^2) - (\sum x_1x_2)^2]$$
- Thus, $b_1 = [(194.875)(1162.5) - (-200.375)(-953.5)] / [(263.875)(194.875) - (-200.375)^2]$
= **3.148**
- The formula to calculate b_2 is:
$$[(\sum x_1^2)(\sum x_2y) - (\sum x_1x_2)(\sum x_1y)] / [(\sum x_1^2)(\sum x_2^2) - (\sum x_1x_2)^2]$$
- Thus, $b_2 = [(263.875)(-953.5) - (-200.375)(1152.5)] / [(263.875)(194.875) - (-200.375)^2]$
= **-1.656**
- The formula to calculate b_0 is: $y - b_1X_1 - b_2X_2$
- Thus, $b_0 = 181.5 - 3.148(69.375) - (-1.656)(18.125)$
= **-6.867**

- Step 4: Place b_0 , b_1 , and b_2 in the estimated linear regression equation.
- The estimated linear regression equation is:
 - $\hat{y} = b_0 + b_1 * x_1 + b_2 * x_2$
 - $\hat{y} = -6.867 + 3.148x_1 - 1.656 x_2$
 - Example of prediction:
 - Suppose Value $x_1= 76$ and $x_2=13$ then value of y ?
 - $= -6.867 + 3.148 * 76 - 1.656 * 13$
 - $=239.248-21.528-6.867$
 - $239.248-28.395=210.853$
 - Additionally,
 - $-6.867 + 3.148 * 70 - 1.656 * 20$
 - $=180.373$

Thanks!!!!

Lecture 9

Graph for Frequency Distributions

31/01/2022

Frequency Distribution

- One method for simplifying and organizing data is to construct a **frequency distribution**
- A **frequency distribution** is an organized tabulation showing exactly how many individuals are located in each category on the scale of measurement.
- A frequency distribution presents an organized picture of the entire set of scores, and it shows where each individual is located relative to others in the distribution

Frequency Distribution

- Frequency Distribution occurs everywhere in our lives. Meteorological department, Data Scientists, Civil Engineers almost all the professions use frequency distributions in their professions.
- These distributions allow us to get insights from any data, see the trends, and predict the next values or the direction in which the data will go.
- There are two types of frequency distributions
 - grouped and
 - ungrouped.
- Their usage depends on the data on which we are working.
- Their analysis is a really important part of probability and statistics.

Frequency Distributions

- Frequency distributions tell us how frequencies are distributed over the values.
- That is how many values lie between different intervals.
- They give us an idea about the range where most of the values fall and the ranges where values are scarce.
- *A frequency distribution is an overview of all values of some variable and the number of times they occur.*

Frequency Distribution Graphs

- In a **frequency distribution graph**, the score categories (X values) are listed on the X axis and the frequencies are listed on the Y axis.
- When the score categories consist of numerical scores from an interval or ratio scale, the graph should be either a histogram or a polygon.

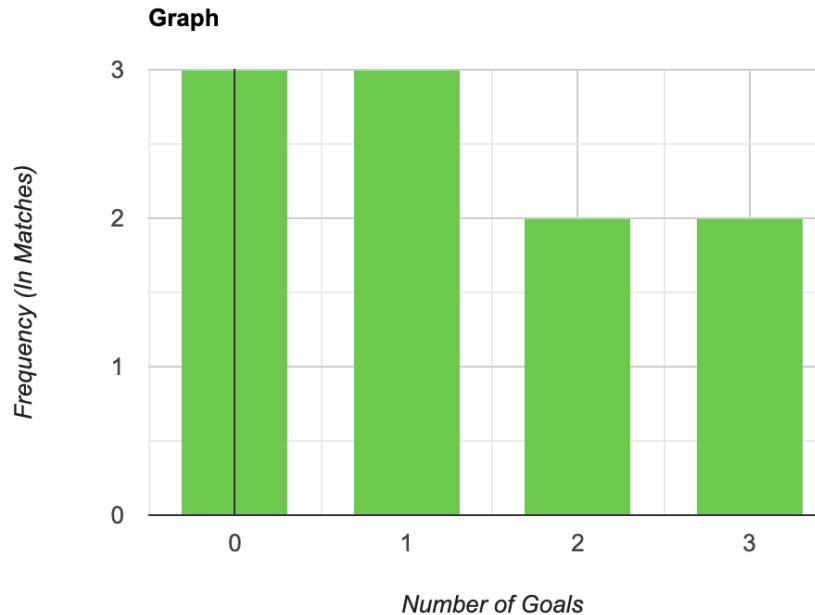
Types Frequency Distributions

- 1. Grouped Frequency Distributions-** Values are divided between different intervals and then their frequencies are counted.
- 2. Un-Grouped Frequency Distributions-** All distinct values of the variable are mentioned and their frequencies are counted.

Example 1

- In a football matched the number of the goals scored by a team in 10 different matches are as follows:
 - 1, 0, 0, 3, 2, 0, 2, 3, 1, 1
- Draw a frequency table to represent this data.

Number of Goals	Frequency
0	3
1	3
2	2
3	2
Total	10



What is Histogram?

- A **histogram** is a graphical representation of a grouped frequency distribution with continuous classes.
- It is an area diagram and can be defined as a set of rectangles with bases along with the intervals between class boundaries and with areas proportional to frequencies in the corresponding classes.
- In such representations, all the rectangles are adjacent since the base covers the intervals between class boundaries.
- The heights of rectangles are proportional to corresponding frequencies of similar classes and for different classes, the heights will be proportional to corresponding frequency densities.
- In other words, a histogram is a diagram involving rectangles whose area is proportional to the frequency of a variable and width is equal to the class interval.

How to Plot Histogram?

You need to follow the below steps to construct a histogram.

1. Begin by marking the class intervals on the X-axis and frequencies on the Y-axis.
 2. The scales for both the axes have to be the same.
 3. Class intervals need to be exclusive.
 4. Draw rectangles with bases as class intervals and corresponding frequencies as heights.
 5. A rectangle is built on each class interval since the class limits are marked on the horizontal axis, and the frequencies are indicated on the vertical axis.
 6. The height of each rectangle is proportional to the corresponding class frequency if the intervals are equal.
 7. The area of every individual rectangle is proportional to the corresponding class frequency if the intervals are unequal.
- Although histograms seem similar to graphs, there is a slight difference between them. The histogram does not involve any gaps between the two successive bars.

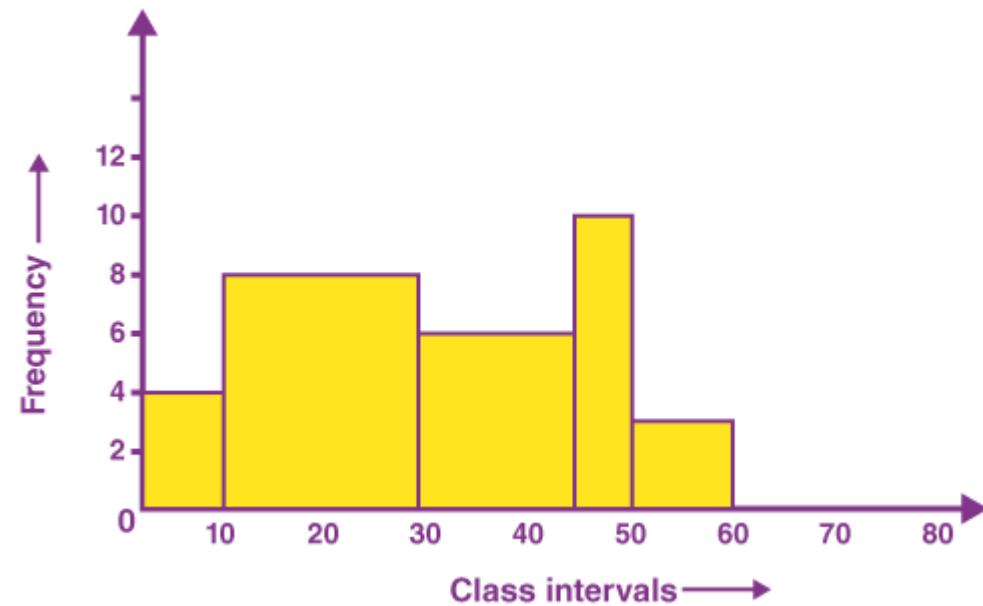
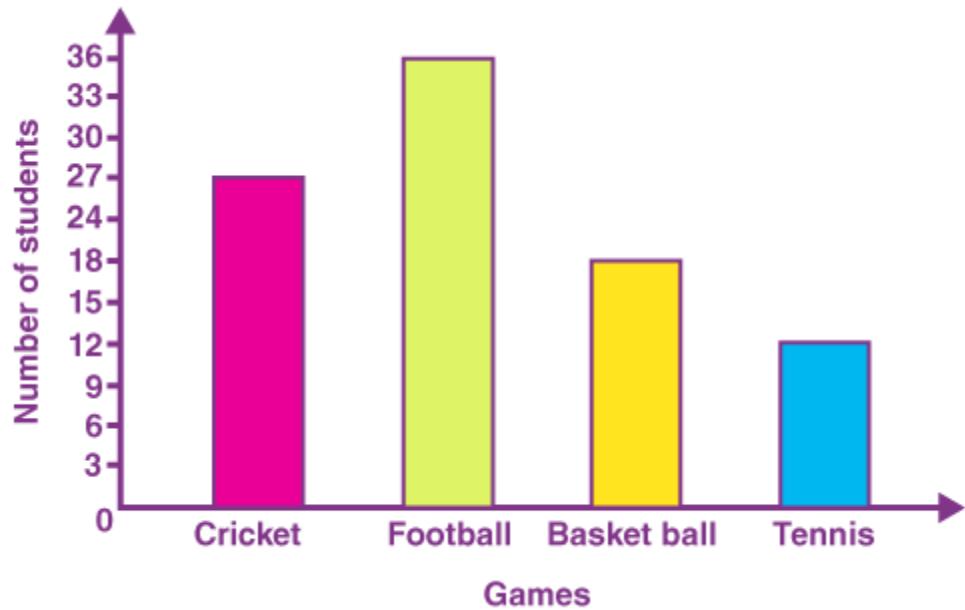
When to Use Histogram?

- The histogram graph is used under certain conditions. They are:
- The data should be numerical.
- A histogram is used to check the shape of the data distribution.
- Used to check whether the process changes from one period to another.
- Used to determine whether the output is different when it involves two or more processes.
- Used to analyse whether the given process meets the customer requirements

Difference Between Bar Graph and Histogram

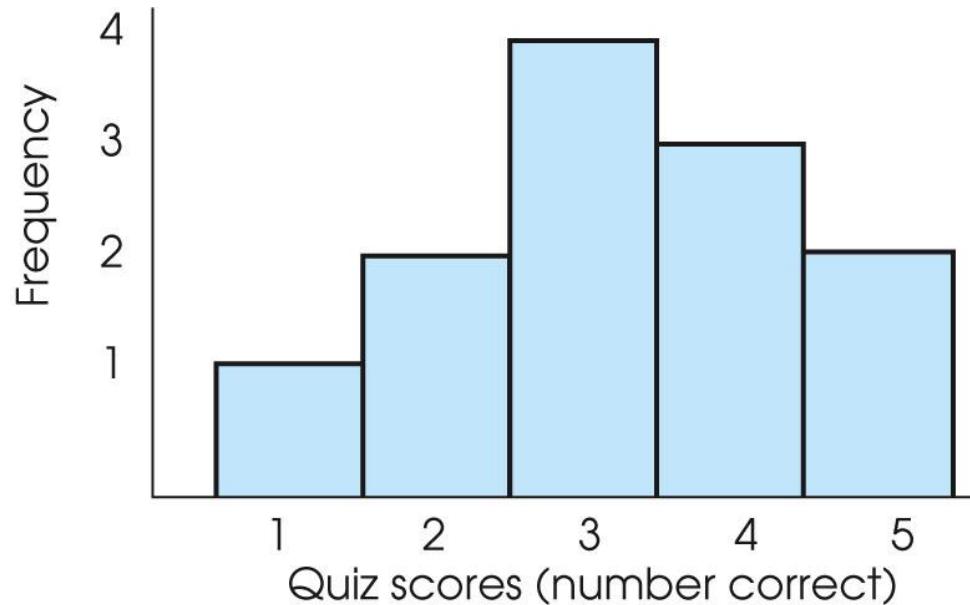
Histogram	Bar Graph
It is a two-dimensional figure	It is a one-dimensional figure
The frequency is shown by the area of each rectangle	The height shows the frequency and the width has no significance.
It shows rectangles touching each other	It consists of rectangles separated from each other with equal spaces.

Difference Between Bar Graph and Histogram

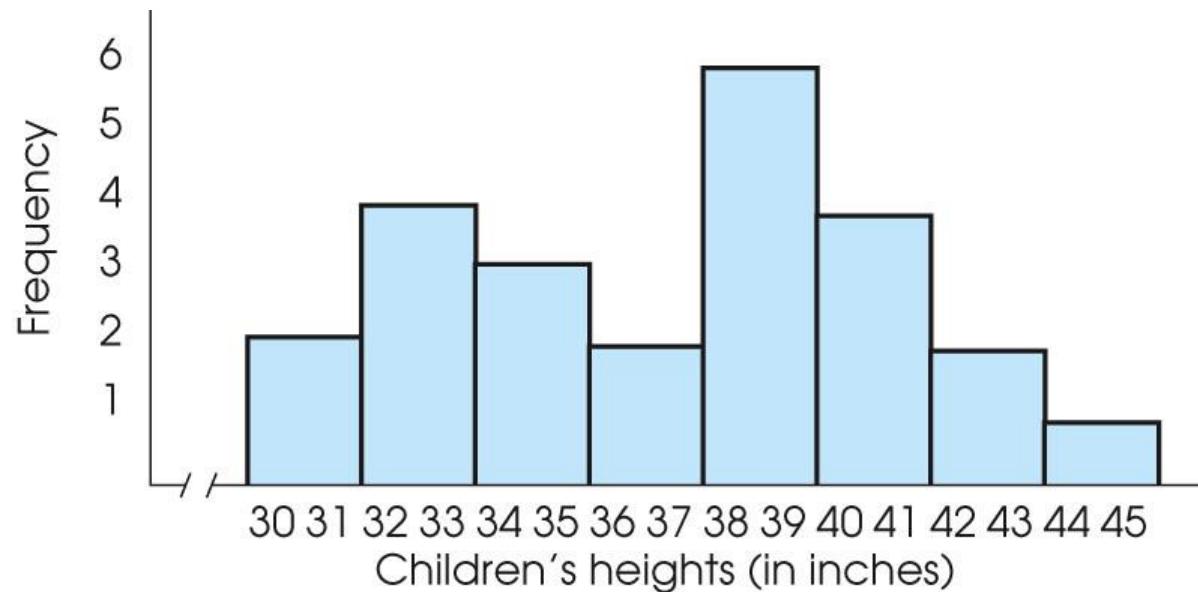


Histograms

- In a **histogram**, a bar is centered above each score (or class interval) so that the height of the bar corresponds to the frequency and the width extends to the real limits, so that adjacent bars touch.



X	f
5	2
4	3
3	4
2	2
1	1

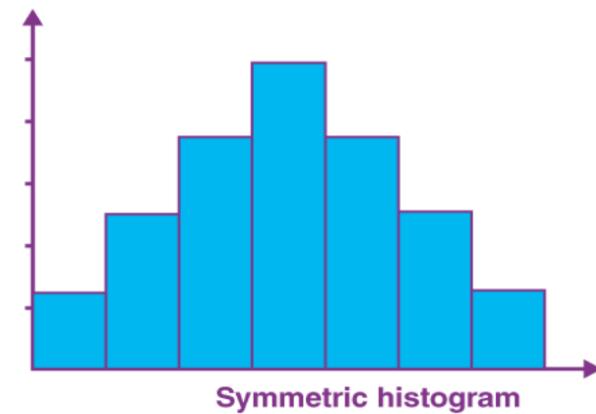
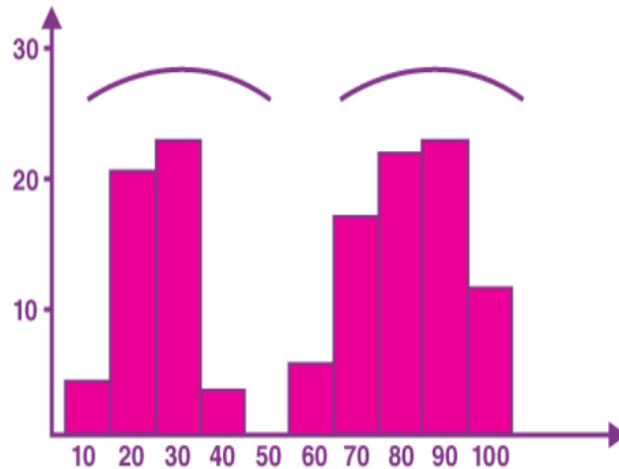
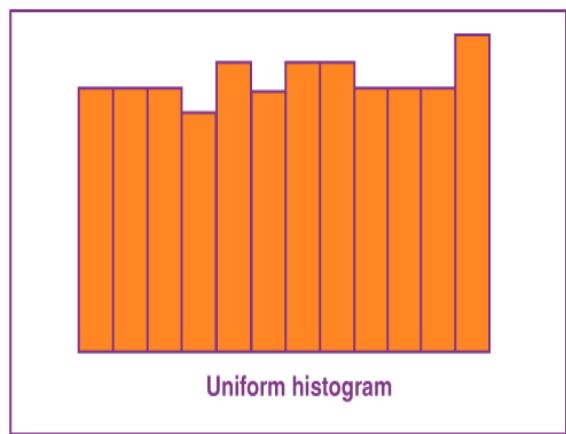


X	f
44–45	1
42–43	2
40–41	4
38–39	6
36–37	2
34–35	3
32–33	4
30–31	2

Kinds of Histogram

- Uniform histogram: A uniform distribution reveals that the number of classes is too small, and each class has the same number of elements. It may involve distribution that has several peaks
- Symmetric histogram: A symmetric histogram is also called a bell-shaped histogram
- Bimodal histogram: If a histogram has two peaks, it is said to be bimodal. Bimodality occurs when the data set has observations on two different kinds of individuals or combined groups if the centers of the two separate histograms are far enough to the variability in both the data sets.
- Probability histogram: A Probability Histogram shows a pictorial representation of a discrete probability distribution. It consists of a rectangle centered on every value of x , and the area of each rectangle is proportional to the probability of the corresponding value.

Types of histogram



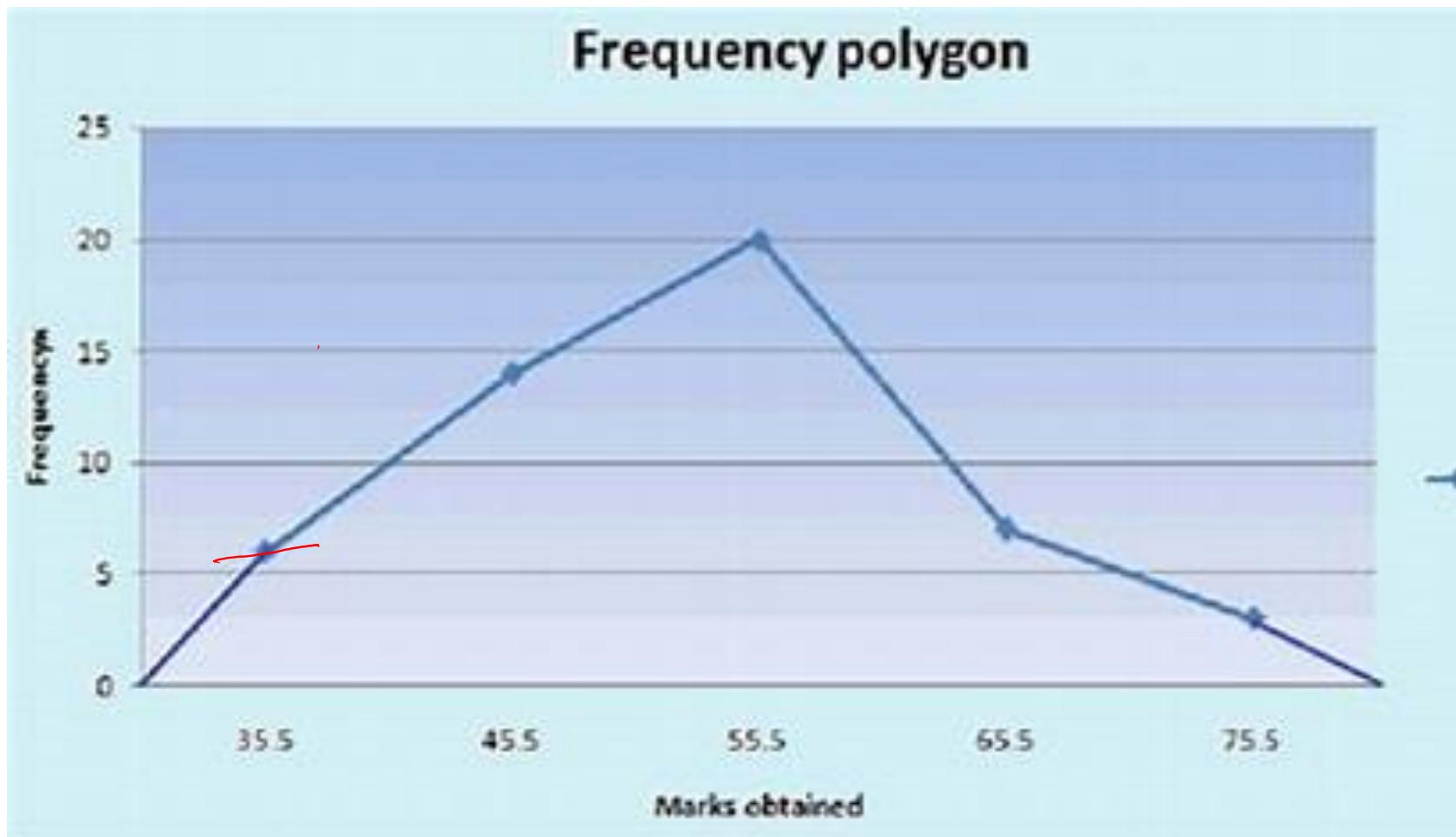
Polygons

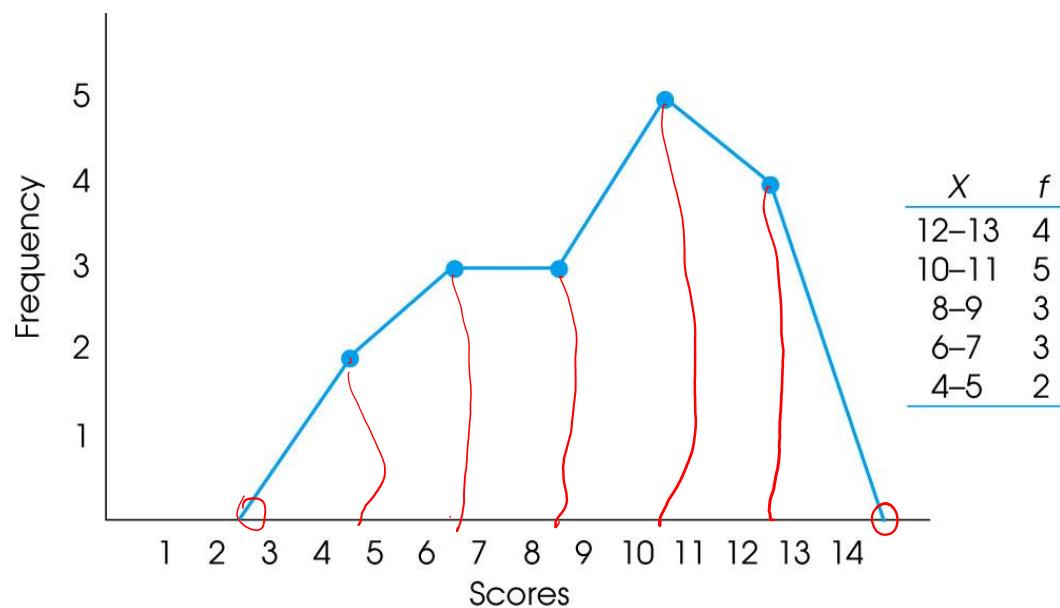
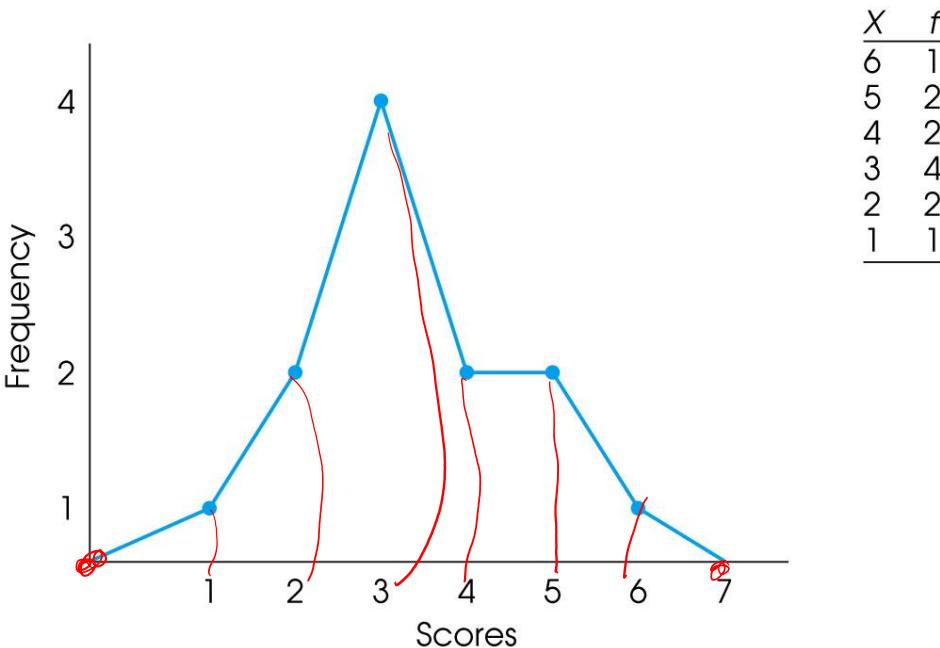
- In a **polygon**, a dot is centered above each score so that the height of the dot corresponds to the frequency. The dots are then connected by straight lines. An additional line is drawn at each end to bring the graph back to a zero frequency.

Frequency Polygon

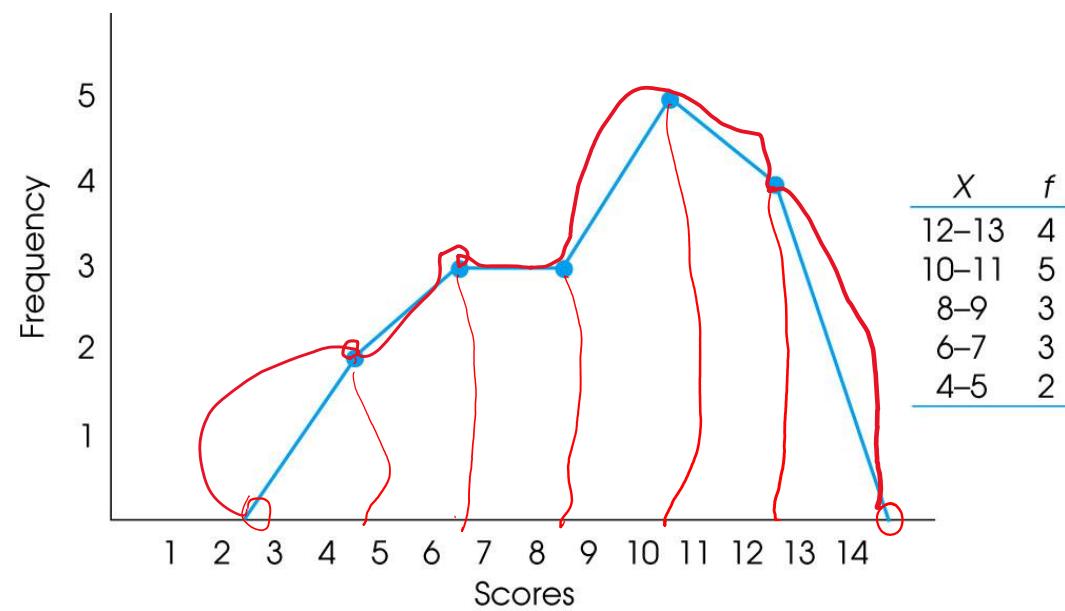
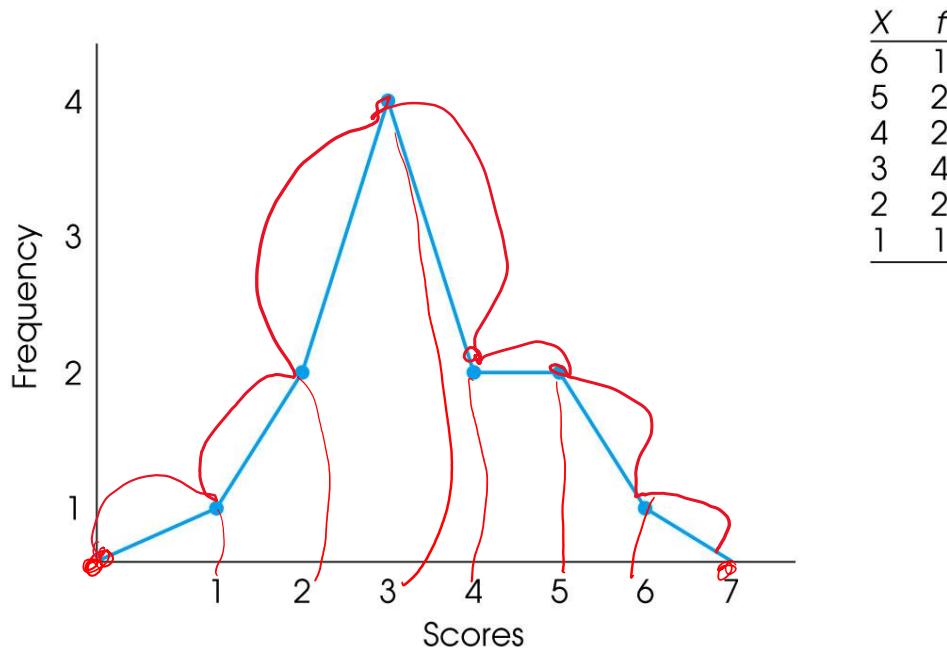
- A frequency polygon is a type of **line graph** where the class frequency is plotted against the class midpoint and the points are joined by a line segment creating a curve. The curve can be drawn with and without a histogram. A frequency polygon graph helps in depicting the highs and lows of frequency distribution data

Frequency Polygon Example



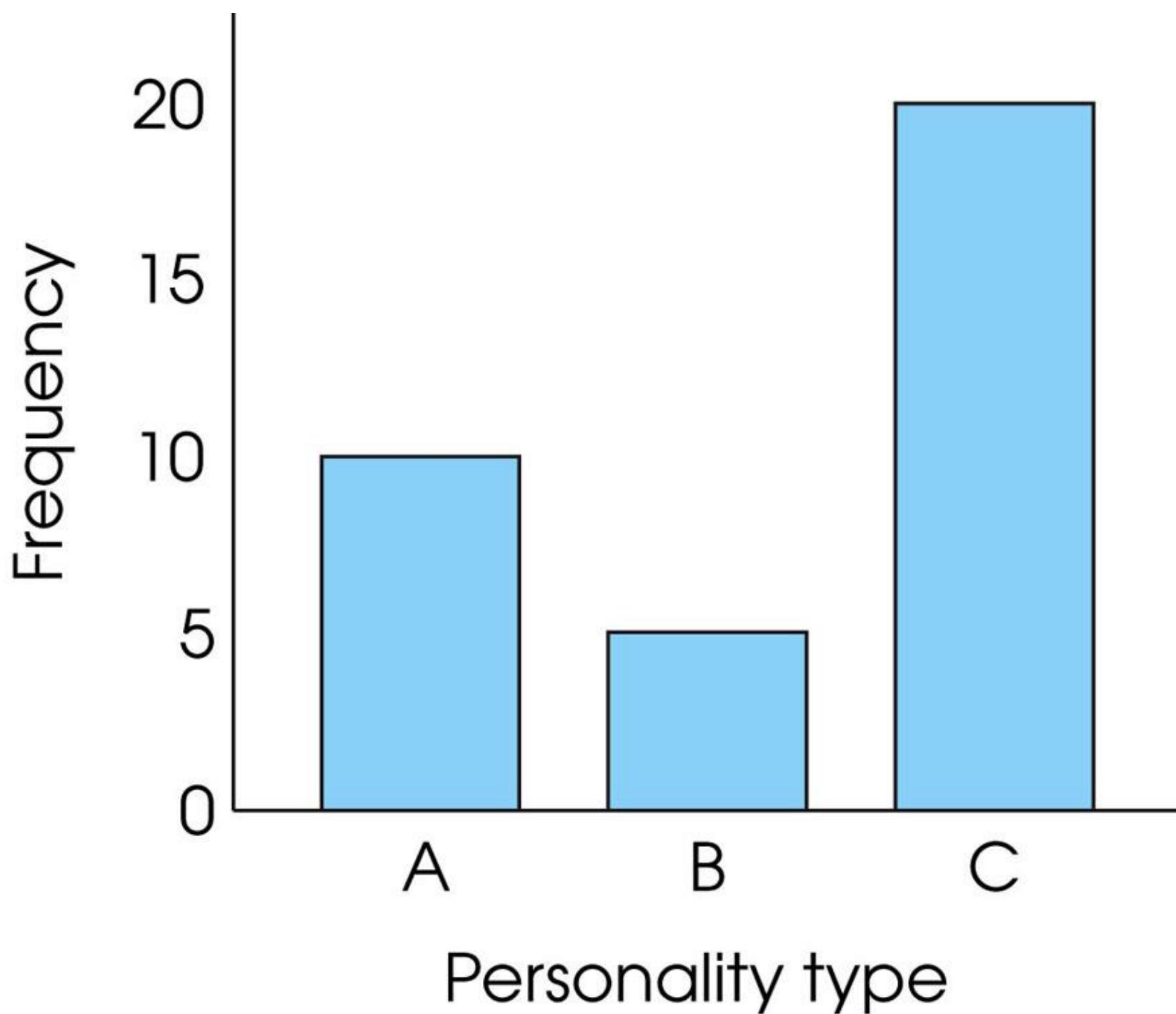


Frequency Curve : Free Hand curve to
draw the graph.



Bar graphs

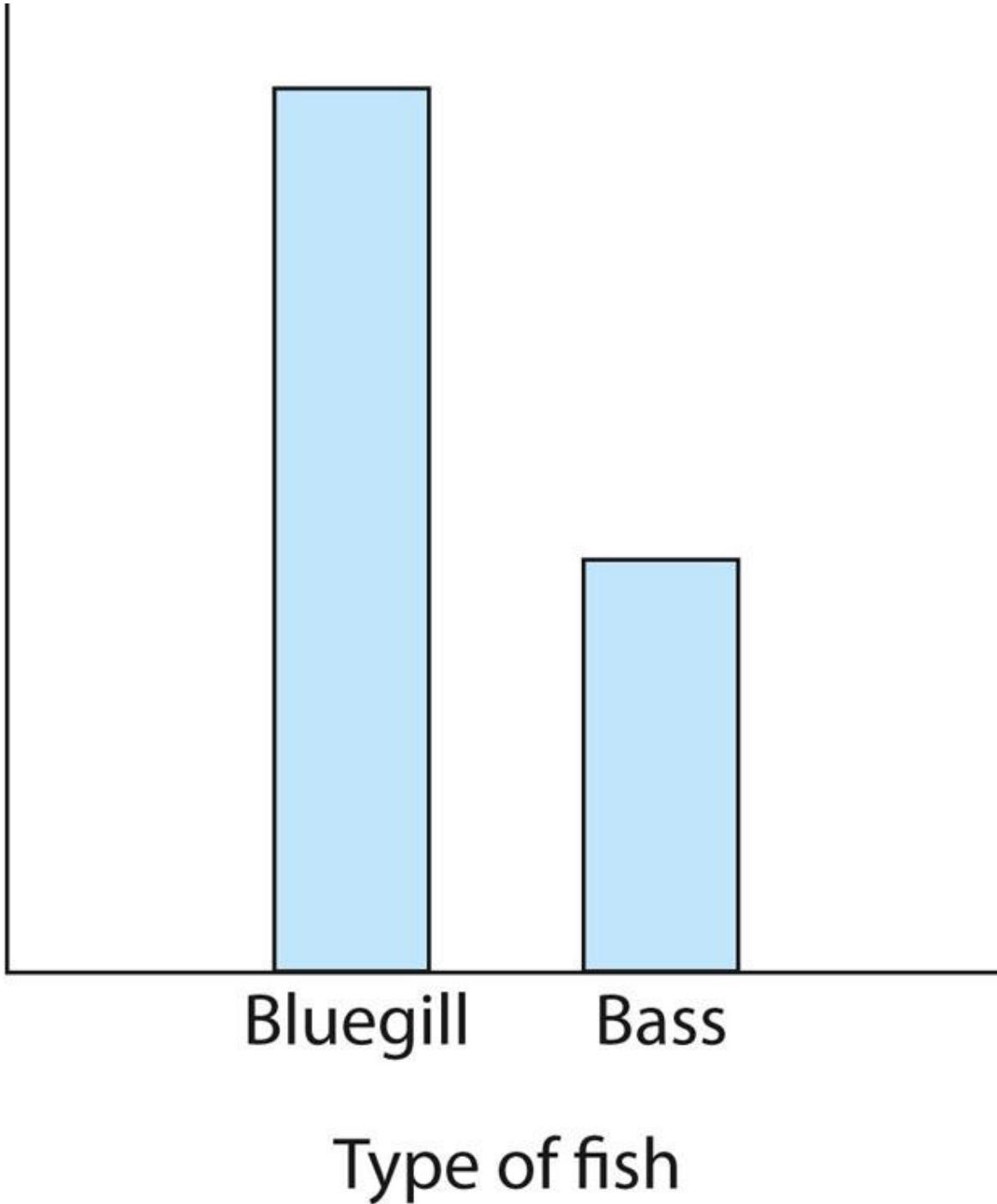
- When the score categories (X values) are measurements from a nominal or an ordinal scale, the graph should be a bar graph.
- A **bar graph** is just like a histogram except that gaps or spaces are left between adjacent bars.



Relative frequency

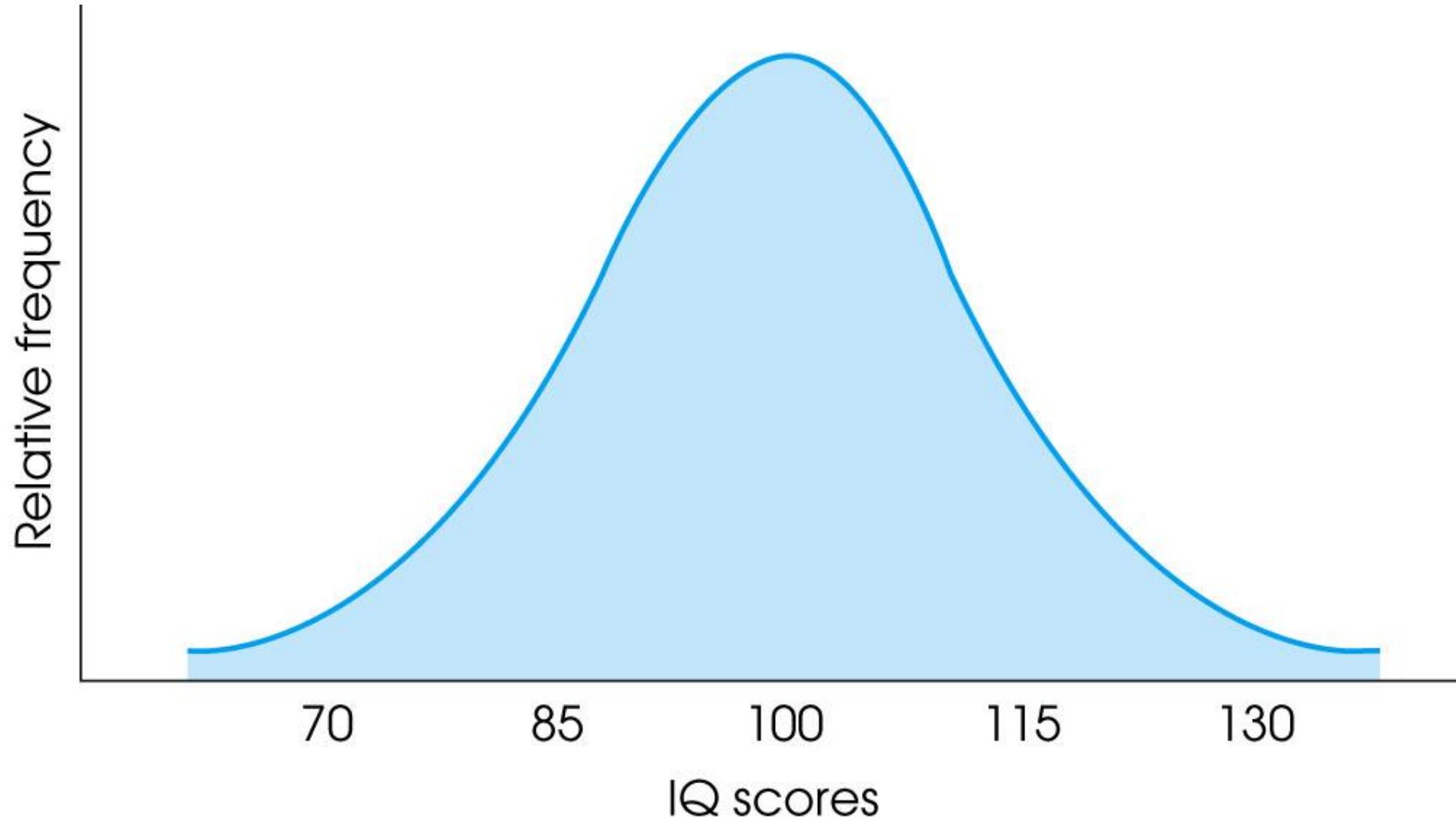
- Many populations are so large that it is impossible to know the exact number of individuals (frequency) for any specific category.
- In these situations, population distributions can be shown using **relative frequency** instead of the absolute number of individuals for each category.

Relative frequency



Smooth curve

- If the scores in the population are measured on an interval or ratio scale, it is customary to present the distribution as a **smooth curve** rather than a jagged histogram or polygon.
- The smooth curve emphasizes the fact that the distribution is not showing the exact frequency for each category.



Frequency distribution graphs

- Frequency distribution graphs are useful because they show the entire set of scores.
- At a glance, you can determine the highest score, the lowest score, and where the scores are centered.
- The graph also shows whether the scores are clustered together or scattered over a wide range.

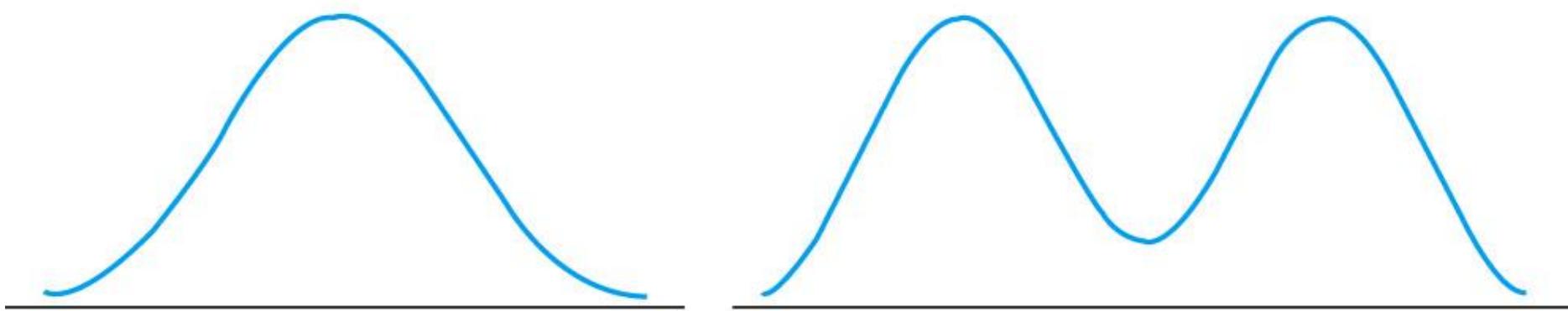
Shape

- A graph shows the **shape** of the distribution.
- A distribution is **symmetrical** if the left side of the graph is (roughly) a mirror image of the right side.
- One example of a symmetrical distribution is the bell-shaped normal distribution.
- On the other hand, distributions are **skewed** when scores pile up on one side of the distribution, leaving a "tail" of a few extreme values on the other side.

Positively and Negatively Skewed Distributions

- In a **positively skewed** distribution, the scores tend to pile up on the left side of the distribution with the tail tapering off to the right.
- In a **negatively skewed** distribution, the scores tend to pile up on the right side and the tail points to the left.

Symmetrical distributions



Skewed distributions



Positive skew

Negative skew

Summary

- Frequency Distribution graph
- Bar graph Vs Histogram
- Types of Histogram and use
- Frequency polygon
- Frequency curve
- Frequency distribution advantages
- Different frequency distribution with its meaning

Lecture 10: Plotting Graphs

2/2/2022

Agenda

- Revision of Some past modules
- Revise Cumulative Frequency
- Cumulative Frequency Curve
- Quantile-Quantile Plot
- Scatter plot

DATA

Facts or figures, which are numerical or otherwise, collected with a definite purpose are called data.

Types Of Data

Quantitative Data

These represent numerical value.

These can be numerically computed.

Qualitative Data

These represent some characteristics or attributes.

These depict descriptions that may be observed but cannot be computed.

Primary Data

Data collected for first time.

Secondary Data

Data that is sourced by someone other than the user.

Discrete Data

These are the data that can take only specific value.

Continuous Data

These are the data that can take values from a given range.

Frequency Distribution Table

A list, table or graph that displays the frequency of various outcomes in a sample of data.

Frequency Distribution Table

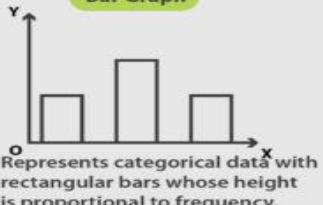
Ungrouped

Marks Obtained	Frequency
16	3
17	4
18	8
19	10
20	12
21	6
22	3

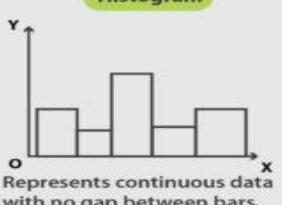
Grouped

Class Interval	Frequency
0-5	3
5-10	11
10-15	14
15-20	2

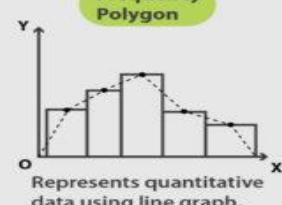
Bar Graph



Histogram



Frequency Polygon



Mean for Ungrouped Data

Let the data set be $x_1, x_2, x_3, \dots, x_n$

$$\text{mean} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n}$$

Mean for Grouped Data

(1) Direct Method

$$\text{mean} = \frac{\sum x_i f_i}{\sum f_i}$$

Where

x_i = Corresponding class mark

f_i = Corresponding frequency

(2) Assumed mean method

$$\text{mean} = a + \frac{\sum d_i f_i}{\sum f_i}$$

Where

a = Assumed mean for the given data

d_i = deviation = $x_i - a$

x_i = Corresponding class mark

f_i = Corresponding frequency

(3) Step Deviation method

$$\text{mean} = a + \frac{\sum f_i u_i}{\sum f_i} \times h$$

Where

a = Assumed mean for the given data

$$u_i = \frac{x_i - a}{h}$$

h = Class width

x_i = Corresponding class mark

f_i = Corresponding frequency

What is Cumulative Frequency?

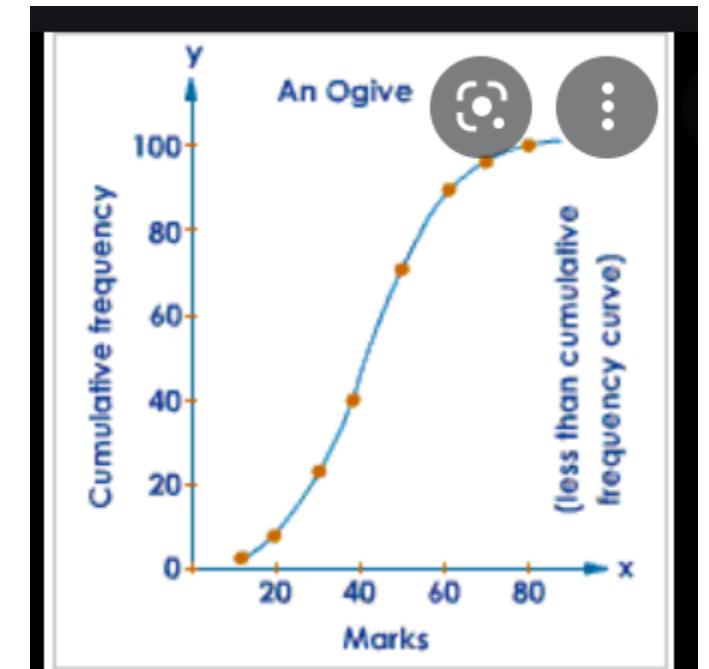
- The frequency is the number of times an event occurs within a given scenario.
- Cumulative frequency is defined as the running total of frequencies.
- It is the sum of all the previous frequencies up to the current point.
- OR we can say, the cumulative frequency of a class is the frequency calculated by adding the frequencies of all the classes preceding the given class.

Example of Cumulative Frequency

Marks	Frequency (No. of Students)	Cumulative Frequency
0 – 5	2	2
5 – 10	10	12
10 – 15	5	17
15 – 20	5	22

What is Cumulative Frequency Distribution?

- A curve that represents the cumulative frequency distribution of grouped data on a graph is called a Cumulative Frequency Curve or an Ogive.
- Representing cumulative frequency data on a graph is the most efficient way to understand the data and derive results.



Types of Cumulative Frequency Curves

- There are two types of Cumulative Frequency Curves (or Ogives) :
 - More than type Cumulative Frequency Curve
 - Less than type Cumulative Frequency Curve

Type-1 : More Than Type Cumulative Frequency Curve

- Here we use the lower limit of the classes to plot the curve.
- How to plot a More than type Ogive:
 - 1.In the graph, put the lower limit on the x-axis
 - 2.Mark the cumulative frequency on the y-axis.
 - 3.Plot the points (x,y) using lower limits (x) and their corresponding Cumulative frequency (y)
 - 4.Join the points by a smooth freehand curve. It looks like an upside down S.

Type -2 Less Than Type Cumulative Frequency Curve

- Here we use the upper limit of the classes to plot the curve.
- How to plot a Less than type Ogive:
 - 1.In the graph, put the upper limit on the x-axis
 - 2.Mark the cumulative frequency on the y-axis.
 - 3.Plot the points (x,y) using upper limits (x) and their corresponding Cumulative frequency (y)
 - 4.Join the points by a smooth freehand curve. It looks like an elongated S.

How to draw Cumulative Graphs?

- Cumulative Graphs can also be used to calculate the Median of given data.
- If you draw both the curves on the same graph, the point at which they intersect, the corresponding value on the x-axis, represents the Median of the given data set.

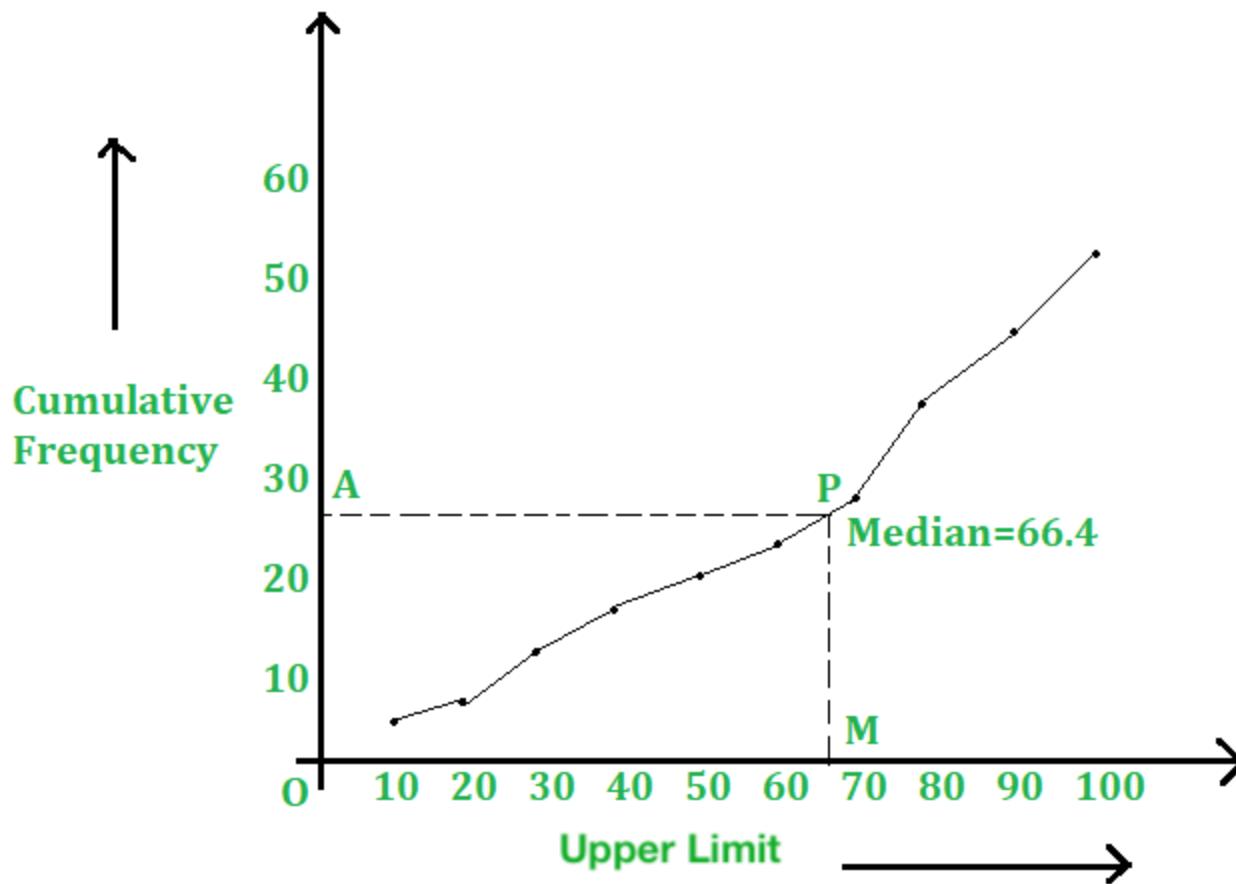
How to draw less than cumulative frequency curve?

In this case, we use the upper limit of the classes to draw the curve.

Now, the step-by-step process of plotting a less than cumulative frequency curve:

1. Take a graph paper and mark the upper-class limits along the x-axis and the corresponding cumulative frequencies along the y-axis.
2. Join these points successively by line segments, we will get a polygon, known as a cumulative frequency polygon.
3. Join these points successively by a smooth curve, we will get a curve, known as cumulative frequency graph.
4. Take a point A (0, N/2) on the y-axis and draw AP || x-axis, cutting the above curve at a point P. Draw PM \perp to the x-axis, cutting the x-axis at M.
5. Then, the median length of OM.

Less than cumulative frequency curve



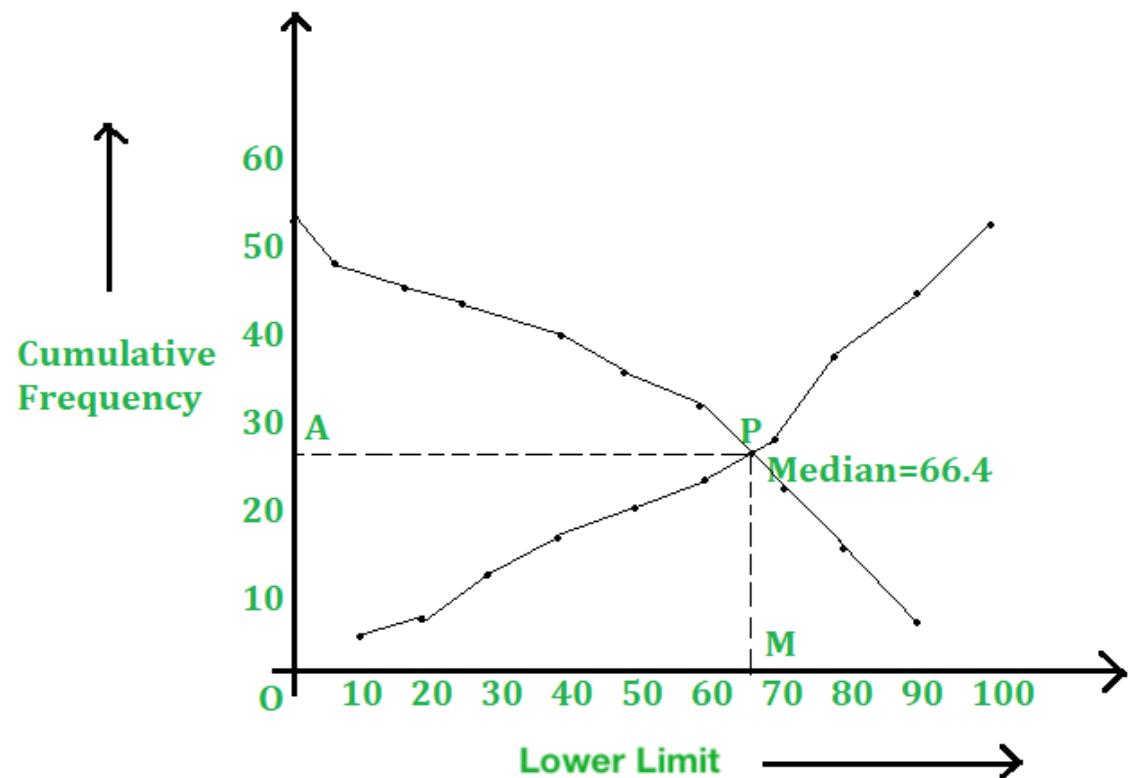
More than cumulative frequency curve

- As we know that the cumulative frequency curves are created using the cumulative frequencies so, in more than cumulative frequency curve, the frequencies of succeeding class or interval are added to the current class or interval frequency.
- You can create more than cumulative frequency by subtracting the frequency of the second-class from the first class and so on.

How to draw more than cumulative frequency curve?

- In this case, we use the lower limit of the classes to draw the curve. Now, the step-by-step process of plotting a more than Cumulative Frequency curve:
 1. Take a graph paper and mark the lower class limits along the x-axis and the corresponding cumulative frequencies along the y-axis.
 2. Join these points successively by line segments, we will get a polygon, known as a cumulative frequency polygon.
 3. Join these points successively by a smooth curve, we will get a curve, known as cumulative frequency graph.
 4. We assume that P be the point of intersection of less than' and 'more than curves. Draw PM \perp to the y-axis, cutting x-axis at M.
 5. Then, median = length of OM.

Example of More than cumulative frequency curve



Example 1:

- Following is the age distribution of group students.
- Now, draw the cumulative frequency curve of less than type and find the median value.

Age (in years)	Frequency
4-5	36
5-6	42
6-7	52
7-8	60
8-9	68
9-10	84
10-11	96
11-12	82
12-13	66
13-14	48
14-15	50
15-16	16

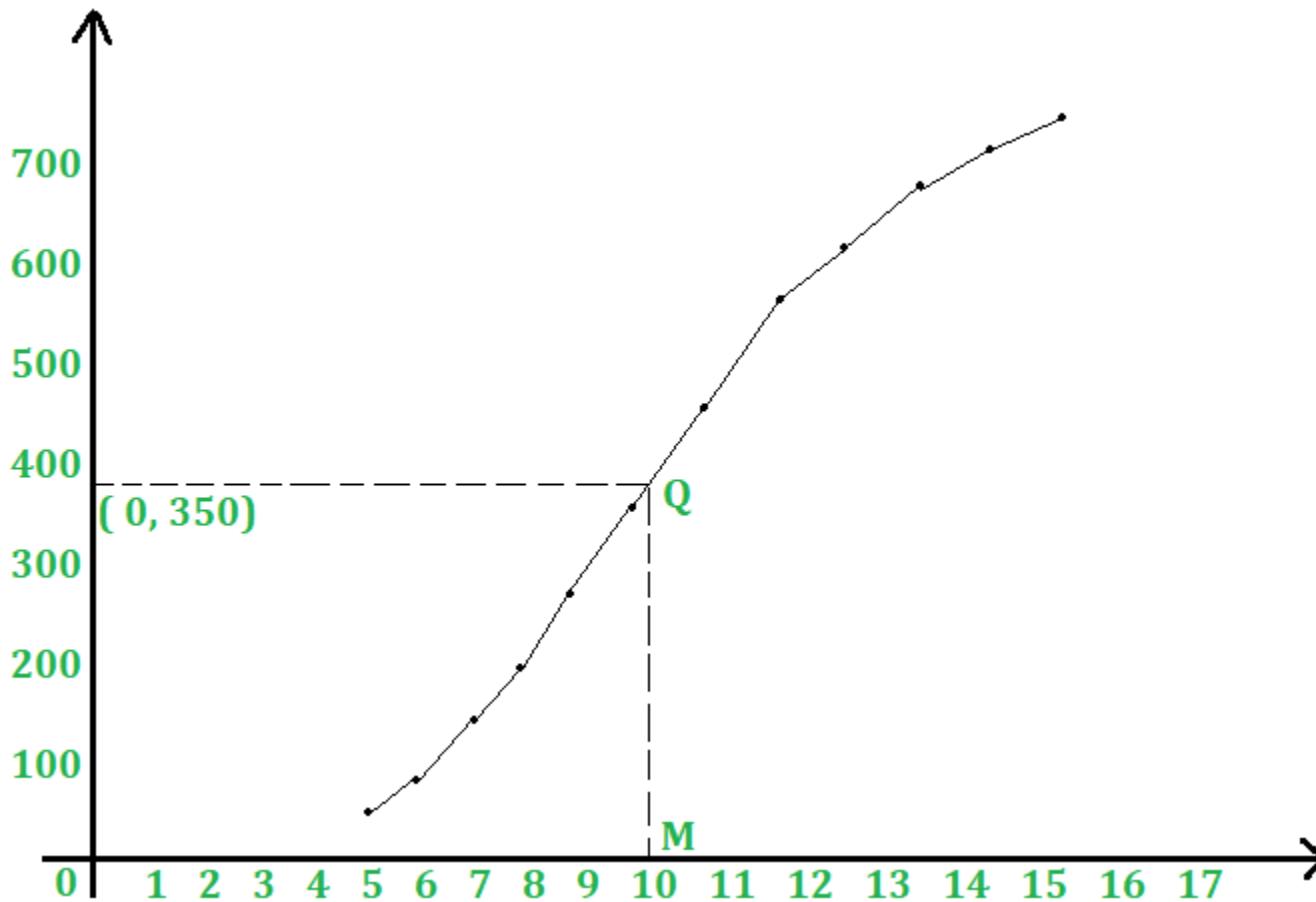
Add CF in Given Table

Age (in years)	Frequency	c.f.
Less than 5	36	36
Less than 6	42	78
Less than 7	52	130
Less than 8	60	190
Less than 9	68	258
Less than 10	84	342
Less than 11	96	438
Less than 12	82	520
Less than 13	66	586
Less than 14	48	634
Less than 15	50	684
Less than 16	16	700

Procedure

- *On a graph paper, take the scale*
- *Along the x-axis: 5 small div. = 1.*
- *Along the y-axis: 1 small div. = 10.*
- *And, plot all the points A(5, 36), B(6, 78), C(7, 130), D(8, 190), E(9, 258), F(10, 342), G(11, 438), H(12, 520), I(13, 586), J(14, 634), K(15, 684) and L(16, 700).*
- *Join these points successively with a freehand, we will get the cumulative frequency curve or an ogive.*
- *Here, $N = 700 \Rightarrow N/2 = 350$*

Ogive Curve



Quantiles and Quantile Based Plots

Quantile–Quantile Plot

- The quantile-quantile plot is a graphical method for determining whether two samples of data came from the same population or not.
- A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.
- By a quantile, we mean the fraction (or percent) of points below the given value.
- For the reference purpose, a 45% line is also plotted, if the samples are from the same population then the points are along this line.

Percentiles and Quantiles

The k -th *percentile* of a set of values divides them so that $k\%$ of the values lie below and $(100 - k)\%$ of the values lie above.

- The 25th percentile is known as the *lower quartile*.
- The 50th percentile is known as the *median*.
- The 75th percentile is known as the *upper quartile*.

It is more common in statistics to refer to *quantiles*. These are the same as percentiles, but are indexed by sample fractions rather than by sample percentages.

Some Difficulties

The previous definition of quantiles and percentiles is not completely satisfactory. For example, consider the six values:

3.7 2.7 3.3 1.3 2.2 3.1

What is the lower quartile of these values?

There is no value which has 25% of these numbers below it and 75% above.

To overcome this difficulty we will use a definition of percentile which is in the spirit of the above statements, but which (necessarily) makes them hold only approximately.

Defining Quantiles

We define the quantiles for the set of values:

3.7 2.7 3.3 1.3 2.2 3.1

as follows.

First sort the values into order:

1.3 2.2 2.7 3.1 3.3 3.7

Associate the ordered values with sample fractions equally spaced from zero to one.

<i>Sample fraction</i>	0	.2	.4	.6	.8	1
<i>Quantile</i>	1.3	2.2	2.7	3.1	3.3	3.7

Defining Quantiles

The other quantiles of

1.3 2.2 2.7 3.1 3.3 3.7

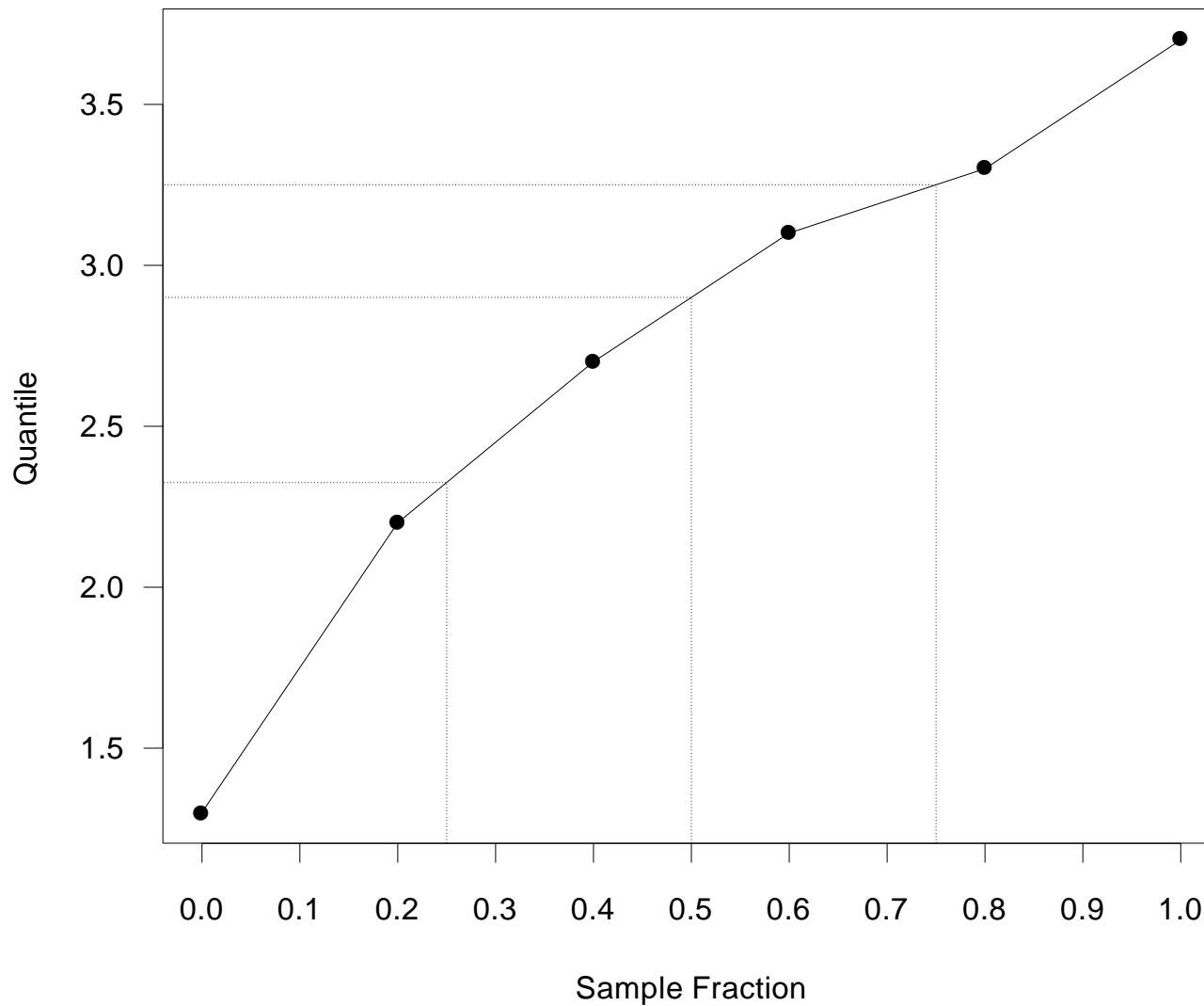
can be obtained by linear interpolation between the values of the table.

The median corresponds to a sample fraction of .5. This lies half way between 0.4 and 0.6. The median must thus be

$$.5 \times 2.7 + .5 \times 3.1 = 2.9$$

The lower quartile corresponds to a sample fraction of .25. This lies one quarter of the way between .2 and .4. The lower quartile must then be $.75 \times 2.2 + .25 \times 2.7 = 2.325$.

Computing the Median and Quartiles



Scatter Plots

- **Scatter plots** are the graphs that present the relationship between two variables in a data-set.
- It represents data points on a two-dimensional plane or on a **Cartesian system**.
- The independent variable or attribute is plotted on the X-axis, while the dependent variable is plotted on the Y-axis.
- These plots are often called **scatter graphs** or **scatter diagrams**.

Scatter plot Graph

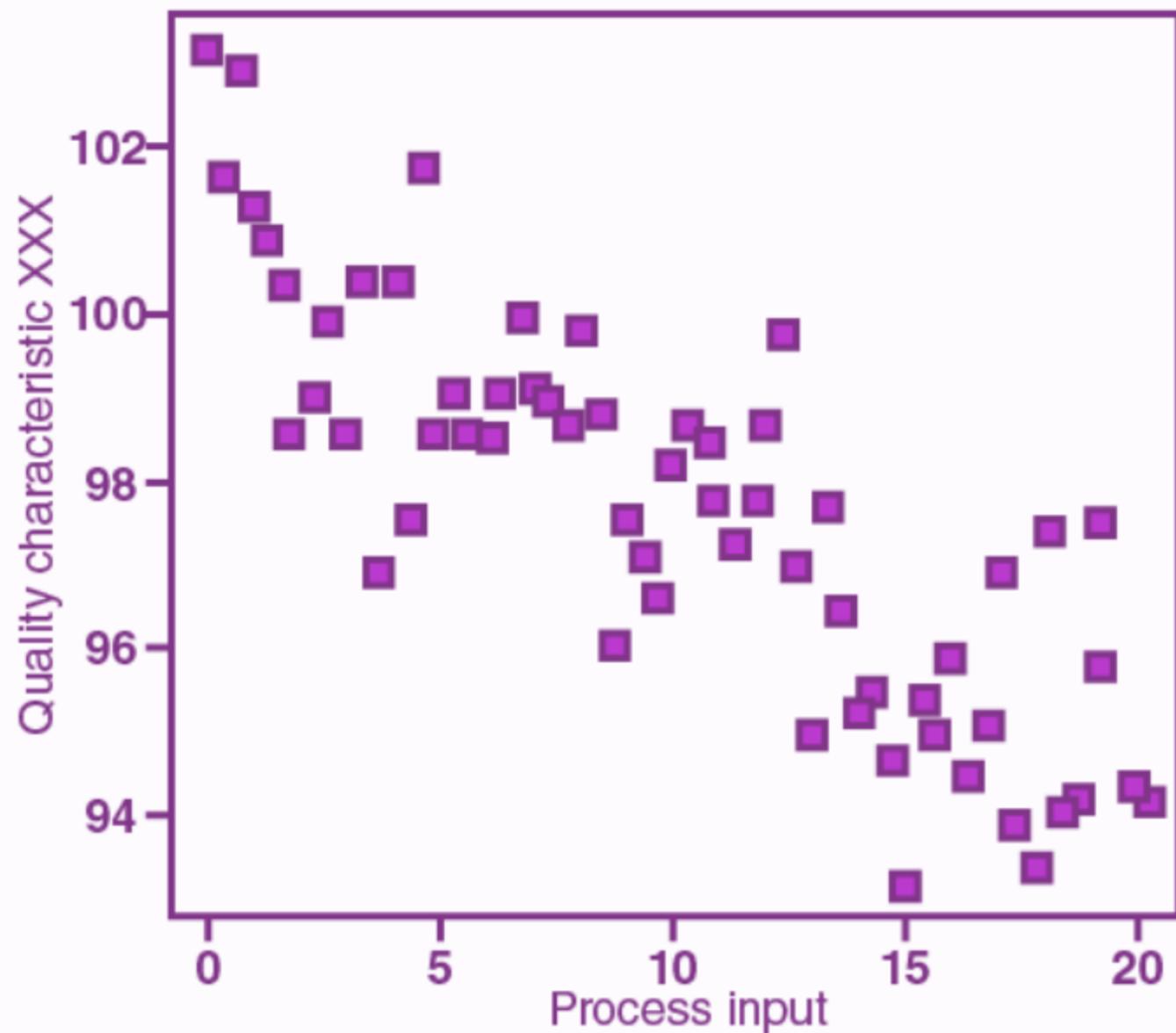
- A scatter plot is also called a scatter chart, scattergram, or scatter plot, XY graph.
- The scatter diagram graphs numerical data pairs, with one variable on each axis, show their relationship.

when to use a scatter plot?

- Scatter plots are used in either of the following situations:
 - ❑ When we have paired numerical data
 - ❑ When there are multiple values of the dependent variable for a unique value of an independent variable
 - ❑ In determining the relationship between variables in some scenarios, such as identifying potential root causes of problems, checking whether two products that appear to be related both occur with the exact cause and so on.

Scatter Plot Uses

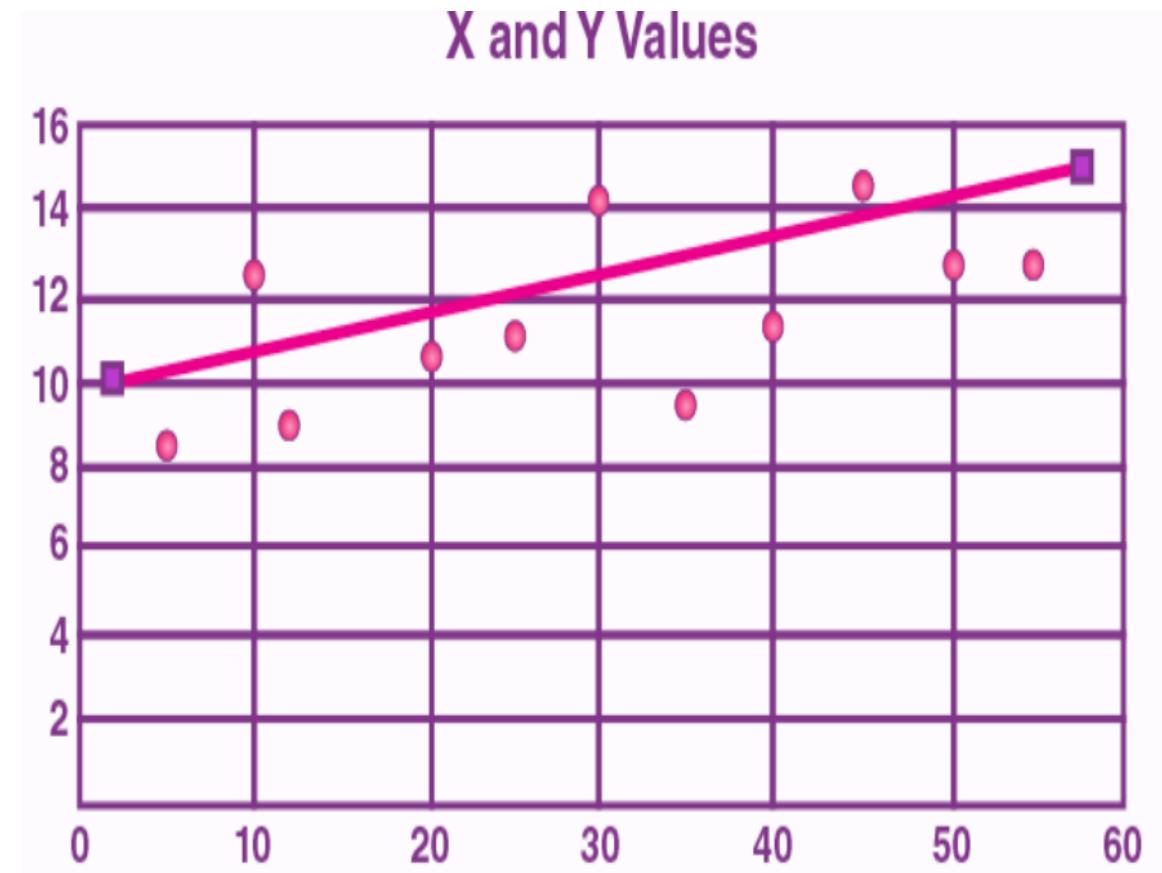
- Scatter plots instantly report a large volume of data. It is beneficial in the following situations –
 - For a large set of data points given
 - Each set comprises a pair of values
 - The given data is in numeric form



Scatterplot for quality characteristic XXX

“line of best fit” or “trend line”

- The line drawn in a scatter plot, which is near to almost all the points in the plot is known as “**line of best fit**” or “**trend line**”



Scatter plot Correlation

- We know that the correlation is a statistical measure of the relationship between the two variables' relative movements. If the variables are correlated, the points will fall along a line or curve. The better the correlation, the closer the points will touch the line. This cause examination tool is considered as one of the seven essential quality tools.

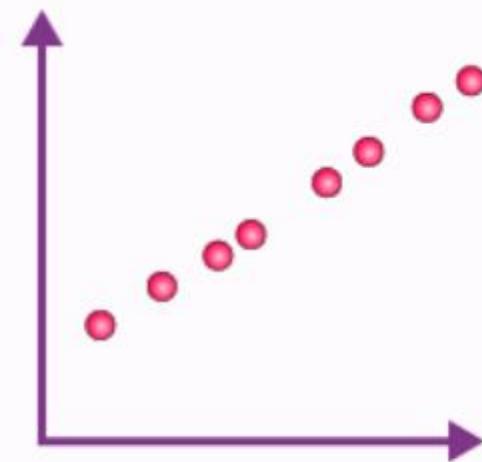
Types of correlation

- The scatter plot explains the correlation between two attributes or variables. It represents how closely the two variables are connected. There can be three such situations to see the relation between the two variables –
 1. Positive Correlation
 2. Negative Correlation
 3. No Correlation

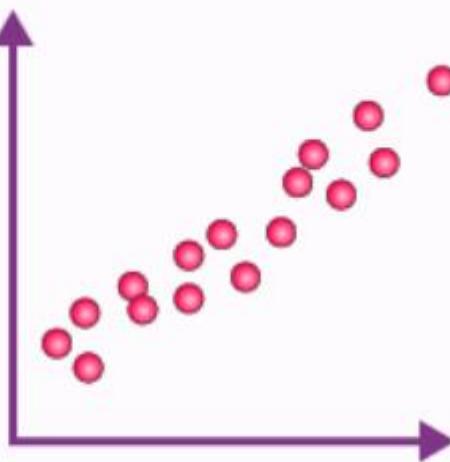
Positive Correlation

- When the points in the graph are rising, moving from left to right, then the scatter plot shows a positive correlation. It means the values of one variable are increasing with respect to another. Now positive correlation can further be classified into three categories:
- **Perfect Positive** – Which represents a perfectly straight line
- **High Positive** – All points are nearby
- **Low Positive** – When all the points are scattered

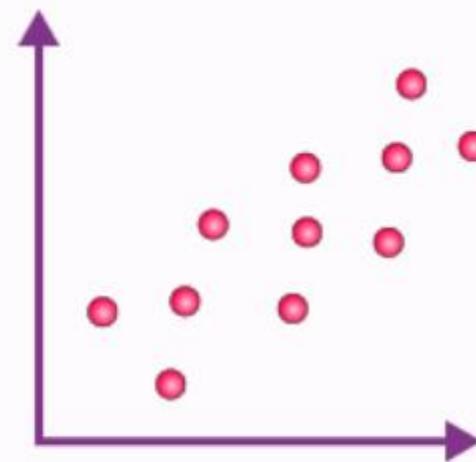
Positive Correlation



Perfect positive correlation



High positive correlation

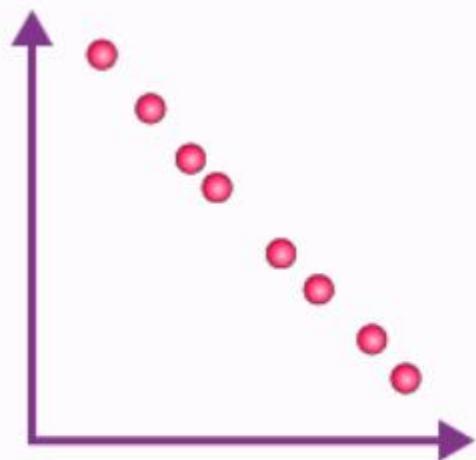


Low positive correlation

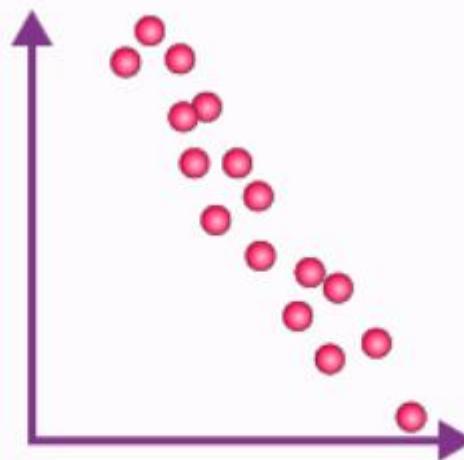
Negative Correlation

- When the points in the scatter graph fall while moving left to right, then it is called a negative correlation. It means the values of one variable are decreasing with respect to another. These are also of three types:
- **Perfect Negative** – Which form almost a straight line
- **High Negative** – When points are near to one another
- **Low Negative** – When points are in scattered form

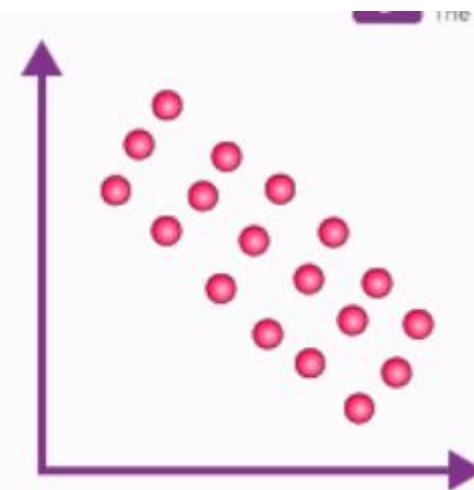
Negative Correlation



Perfect negative correlation



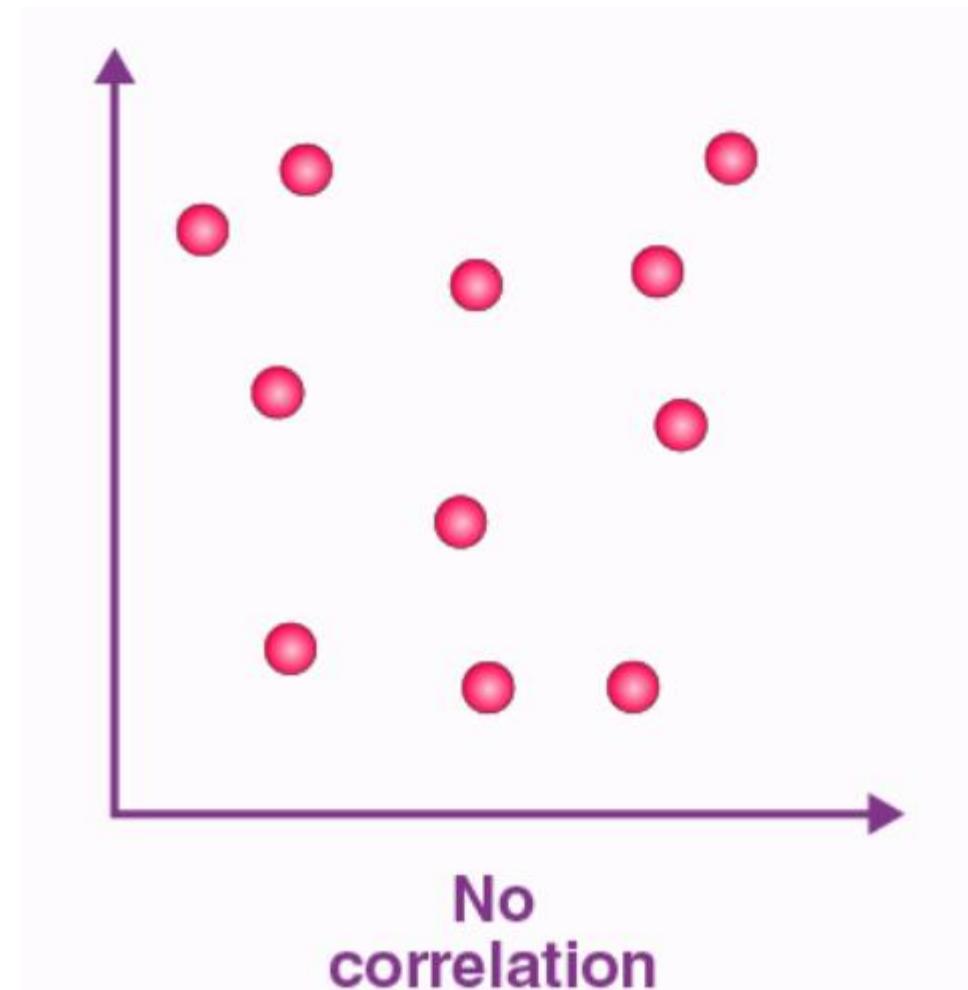
High negative correlation



Low negative correlation

No correlation

When the points are scattered all over the graph and it is difficult to conclude whether the values are increasing or decreasing, then there is no correlation between the variables



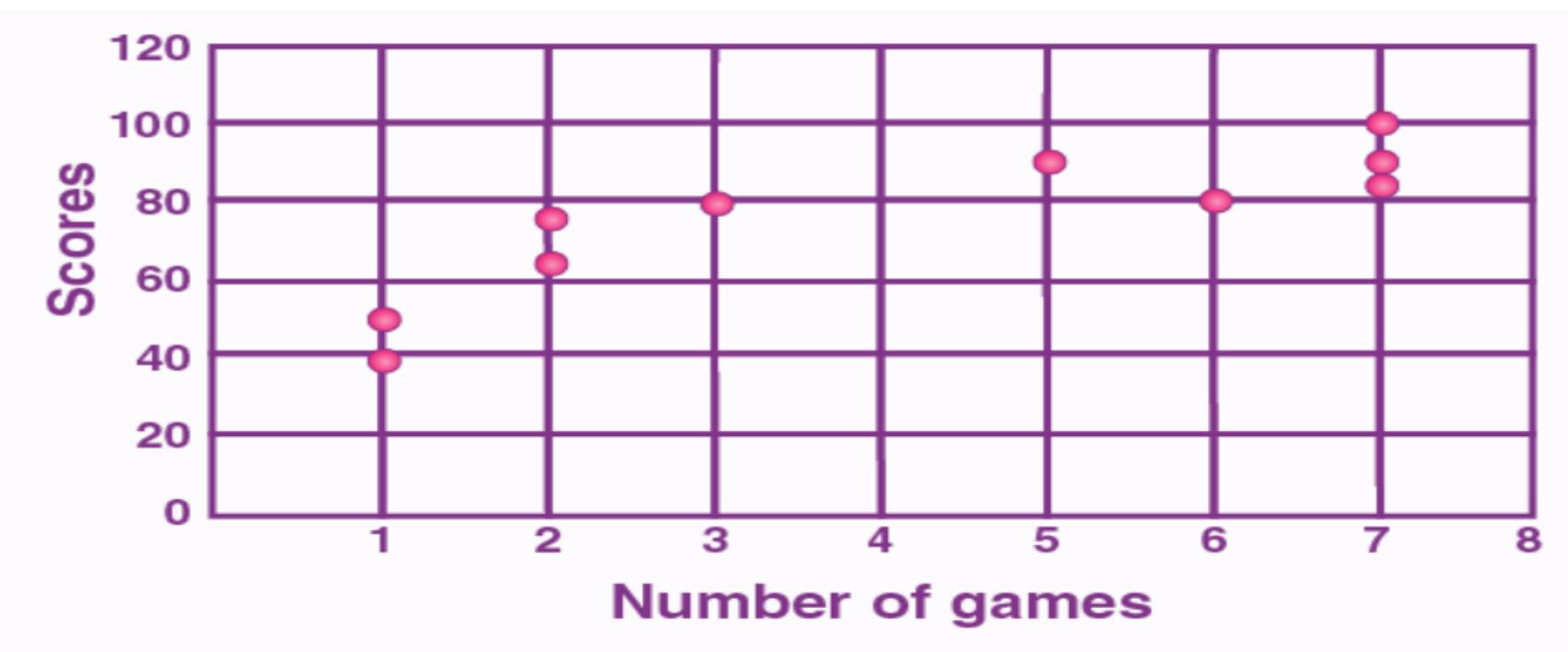
Scatter plot Example

Draw a scatter plot for the given data that shows the number of games played and scores obtained in each instance

No. of games	3	5	2	6	7	1	2	7	1	7
Scores	80	90	75	80	90	50	65	85	40	100

Scatter Plot

- X-axis or horizontal axis: Number of games
- Y-axis or vertical axis: Scores



Summary

- Scatter Plot
- Quantile -Quantile Plot
- Cumulative Frequency Curve

Lecture 11 :Kinds of Deviations

3/02/2022

Deviations

- In mathematic and statistics, **deviation** is a measure of difference between the observed value of a variable and some other value, often that variable's mean.
- The sign of the deviation reports the direction of that difference (the deviation is positive when the observed value exceeds the reference value).
- The magnitude of the value indicates the size of the difference

Deviations

- A deviation that is a difference between an observed value and the *true value* of a quantity of interest (where *true value* denotes the Expected Value, such as the population mean) is an **error**.
- A deviation that is the difference between the observed value and an *estimate* of the true value (e.g. the sample mean)
- The Expected Value of a sample can be used as an estimate of the Expected Value of the population) is a **residual**.
- These concepts are applicable for data at the interval and ratio levels of measurement.

Mean Deviation

- In statistics and mathematics, the deviation is a measure that is used to find the difference between the observed value and the expected value of a variable.
- In simple words, the deviation is the distance from the center point. Similarly, the mean deviation is used to calculate how far the values fall from the middle of the data set.
- In this lecture , let us discuss the definition, formula, and examples in detail.

Mean Deviation Definition

- The mean deviation is defined as a statistical measure that is used to calculate the average deviation from the mean value of the given data set. The mean deviation of the data values can be easily calculated using the below procedure.
- Step 1: Find the mean value for the given data values
- Step 2: Now, subtract the mean value from each of the data values given (Note: Ignore the minus symbol)
- Step 3: Now, find the mean of those values obtained in step 2.

Mean Deviation Formula

- The formula to calculate the mean deviation for the given data set is given below.
- Mean Deviation = $[\Sigma |X - \mu|]/N$
- Here,
- Σ represents the addition of values
- X represents each value in the data set
- μ represents the mean of the data set
- N represents the number of data values
- $| |$ represents the absolute value, which ignores the “-” symbol

Mean Deviation for Frequency Distribution

- To present the data in the more compressed form we group it and mention the frequency distribution of each such group. These groups are known as class intervals.
- Grouping of data is possible in two ways:
 1. Discrete Frequency Distribution
 2. Continuous Frequency Distribution
- In the upcoming discussion, we will be discussing mean absolute deviation in a discrete frequency distribution.
- Let us first know what is actually meant by the discrete distribution of frequency.

Mean Deviation for Discrete Distribution Frequency

- As the name itself suggests, by discrete we mean distinct or non-continuous. In such a distribution the frequency (number of observations) given in the set of data is discrete in nature.
- If the data set consists of values $x_1, x_2, x_3, \dots, x_n$ each occurring with a frequency of f_1, f_2, \dots, f_n respectively then such a representation of data is known as the discrete distribution of frequency.

Calculate Mean Deviation

- To calculate the mean deviation for grouped data and particularly for discrete distribution data the following steps are followed:
- **Step I:** The measure of central tendency about which mean deviation is to be found out is calculated. Let this measure be a .
- If this measure is mean then it is calculated as,

- where $N = \sum_{i=1}^n f_i$

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

$$\Rightarrow \bar{x} = \frac{1}{N} \sum_{i=1}^n x_i f_i$$

Formula

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

$$\Rightarrow \bar{x} = \frac{1}{N} \sum_{i=1}^n x_i f_i$$

Calculation of Mean deviation

- If the measure is median then the given set of data is arranged in ascending order and then the cumulative frequency is calculated then the observations whose cumulative frequency is equal to or just greater than $N/2$ is taken as the [median](#) for the given discrete distribution of frequency and it is seen that this value lies in the middle of the frequency distribution.
- **Step II:** Calculate the absolute deviation of each observation from the measure of central tendency calculated in step (I)
- **Step III:** The mean absolute deviation around the measure of central tendency is then calculated by using the formula

Formula

$$M.A.D(a) = \frac{\sum_{i=1}^n f_i |x_i - a|}{N}$$

If the central tendency is mean then

$$M.A.D(\bar{x}) = \frac{\sum_{i=1}^n f_i |x_i - \bar{x}|}{N}$$

In case of median

$$M.A.D(M) = \frac{\sum_{i=1}^n f_i |x_i - M|}{N}$$

Mean Deviation Examples

- Determine the mean deviation for the data values 5, 3, 7, 8, 4, 9.
- **Solution:**
- Given data values are 5, 3, 7, 8, 4, 9.
- We know that the procedure to calculate the mean deviation.
- First, find the mean for the given data:
- Mean, $\mu = (5+3+7+8+4+9)/6$
- $\mu = 36/6$
- $\mu = 6$
- Therefore, the mean value is 6.

Step 2: Calculate Mean Deviation

- Now, subtract each mean from the data value, and ignore the minus symbol if any
- (Ignore"-")
 - $5 - 6 = 1$
 - $3 - 6 = 3$
 - $7 - 6 = 1$
 - $8 - 6 = 2$
 - $4 - 6 = 2$
 - $9 - 6 = 3$
- Now, the obtained data set is 1, 3, 1, 2, 2, 3.

Final Step to Compute Mean Deviation

- Finally, find the mean value for the obtained data set
- Therefore, the mean deviation is
- $= (1+3 + 1+ 2+ 2+3) /6$
- $= 12/6$
- $= 2$
- Hence, the mean deviation for 5, 3, 7, 8, 4, 9 is 2.

Example 2

- In a foreign language class, there are 4 languages, and the frequencies of students learning the language and the frequency of lectures per week are given below:

Language	Sanskrit	Spanish	French	English
No. of students(x_i)	6	5	9	12
Frequency of lectures(f_i)	5	7	4	9

Calculate the mean deviation about the mean for the given data?

Solution

x_i	f_i	$x_i f_i$	$ x_i - \bar{x} $	$f_i x_i - \bar{x} $
6	5	30	2.36	11.8
5	7	35	3.36	23.52
9	4	36	0.64	2.56
12	9	108	3.64	32.76
	$\sum f_i = 25$	$\bar{x} = \frac{1}{N} \sum_{i=1}^n x_i f_i = 8.36$		$\sum_{i=1}^n f_i x_i - \bar{x} = 70.64$

Quartile Deviations

- Quartile deviation is one of the measures of dispersion.
- The Quartile Deviation is a simple way to **estimate the spread of a distribution about a measure of its central tendency** (usually the mean). So, it gives you an idea about the range within which the central 50% of your sample data lies.
- The Quartile Deviation can be defined mathematically as **half of the difference between the upper and lower quartile**.
- Here, quartile deviation can be represented as QD; Q 3 denotes the upper quartile and Q 1 indicates the lower quartile.
- Quartile Deviation is also known as the Semi Interquartile range.

Quartile Deviation Formula

- Suppose Q_1 is the lower quartile, Q_2 is the median, and Q_3 is the upper quartile for the given data set, then its quartile deviation can be calculated using the following formula.
- $QD = (Q_3 - Q_1)/2$
- In the next section, you will learn how to calculate these quartiles for both ungrouped and grouped data separately.

Quartile Deviation for Ungrouped Data

- For an ungrouped data, quartiles can be obtained using the following formulas,
- $Q_1 = [(n+1)/4]\text{th item}$
- $Q_2 = [(n+1)/2]\text{th item}$
- $Q_3 = [3(n+1)/4]\text{th item}$
- Where n represents the total number of observations in the given data set.
- Also, Q_2 is the median of the given data set, Q_1 is the median of the lower half of the data set and Q_3 is the median of the upper half of the data set.
- Before, estimating the quartiles, we have to arrange the given data values in ascending order. If the value of n is even, we can follow the similar procedure of finding the median.

Quartile Deviation for Grouped Data

$$Q_r = l_1 + \frac{r\left(\frac{N}{4}\right) - c}{f} (l_2 - l_1)$$

Q_r = the rth quartile

l_1 = the lower limit of the quartile class

l_2 = the upper limit of the quartile class

f = the frequency of the quartile class

c = the cumulative frequency of the class preceding the quartile class

N = Number of observations in the given data set

Quartile Deviation Example

- Find the quartiles and quartile deviation of the following data:
- 17, 2, 7, 27, 15, 5, 14, 8, 10, 24, 48, 10, 8, 7, 18, 28

Step 1:

- Given data:
- 17, 2, 7, 27, 15, 5, 14, 8, 10, 24, 48, 10, 8, 7, 18, 28
- Ascending order of the given data is:
- 2, 5, 7, 7, 8, 8, 10, 10, 14, 15, 17, 18, 24, 27, 28, 48
- Number of data values = $n = 16$

Step 2: Compute Median of the given data set

- Q_2 = Median of the given data set
- n is even, median = $(1/2) [(n/2)\text{th observation and } (n/2 + 1)\text{th observation}]$
- = $(1/2)[8\text{th observation} + 9\text{th observation}]$
- = $(10 + 14)/2$
- = $24/2$
- = 12
- $Q_2 = 12$

Step 3: Compute lower half of the data

- Now, is:
- 2, 5, 7, 7, 8, 8, 10, 10 (even number of observations)
- $Q_1 = \text{Median of lower half of the data}$
- $= (1/2)[4\text{th observation} + 5\text{th observation}]$
- $= (7 + 8)/2$
- $= 15/2$
- $= 7.5$

Step 4 : Compute upper half of the data

- Also, the upper half of the data is:
- 14, 15, 17, 18, 24, 27, 28, 48 (even number of observations)
- $Q_3 = \text{Median of upper half of the data}$
- $= (1/2)[4\text{th observation} + 5\text{th observation}]$
- $= (18 + 24)/2$
- $= 42/2$
- $= 21$

Step 5: Find Solution

- Quartile deviation = $(Q_3 - Q_1)/2$
- = $(21 - 7.5)/2$
- = $13.5/2$
- = 6.75
- Therefore, the quartile deviation for the given data set is 6.75.

Example 2: Quartile Deviations

- Calculate the quartile deviation for the following distribution.

Class	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
Frequency	5	3	4	3	3	4	7	9	7	8

Calculate the cumulative frequency for the given distribution of data.

Class	Frequency	Cumulative Frequency
0 – 10	5	5
10 – 20	3	$5 + 3 = 8$
20 – 30	4	$8 + 4 = 12$
30 – 40	3	$12 + 3 = 15$
40 – 50	3	$15 + 3 = 18$
50 – 60	4	$18 + 4 = 22$
60 – 70	7	$22 + 7 = 29$
70 – 80	9	$29 + 9 = 38$
80 – 90	7	$38 + 7 = 45$
90 – 100	8	$45 + 8 = 53$

Formula and its associated value

- Here, N = 53

$$Q_r = l_1 + \frac{r\left(\frac{N}{4}\right) - c}{f} (l_2 - l_1)$$

Finding Q₁:

r = 1

$$N/4 = 53/4 = 13.25$$

Thus, Q1 lies in the interval 30 – 40.

In this case, quartile class = 30 – 40

l_1 = the lower limit of the quartile class = 30

l_2 = the upper limit of the quartile class = 40

f = the frequency of the quartile class = 3

c = the cumulative frequency of the class preceding the quartile class = 12

Step 1 calculating Q1

- Now, by substituting these values in the formula we get:
- $Q_1 = 30 + [(13.25 - 12)/3] \times (40 - 30)$
- $= 30 + (1.25/3) \times 10$
- $= 30 + (12.5/3)$
- $= 30 + 4.167$
- $= 34.167$

Step 2: Finding Q_3 :

- $r = 3$
- $3N/4 = 3 \times 13.25 = 39.75$
- Thus, Q_3 lies in the interval 80 – 90.
- In this case, quartile class = 80 – 90
- l_1 = the lower limit of the quartile class = 80
- l_2 = the upper limit of the quartile class = 90
- f = the frequency of the quartile class = 7
- c = the cumulative frequency of the class preceding the quartile class = 38

Step 3: Calculating Q3

- Now, by substituting these values in the formula we get:
- $Q_3 = 80 + [(39.75 - 38)/7] \times (90 - 80)$
- $= 80 + (1.75/7) \times 10$
- $= 80 + (17.5/7)$
- $= 80 + 2.5$
- $= 82.5$

Final Quartile Deviation

- Finally, the quartile deviation = $(Q_3 - Q_1)/2$
- $QD = (82.5 - 34.167)/2$
- $= 48.333/2$
- $= 24.1665$
- Hence, the quartile deviation of the given distribution is 24.167 (approximately).

Standard Deviation

- **Standard Deviation** is a measure which shows how much variation (such as spread, dispersion, spread,) from the mean exists.
- The standard deviation indicates a “typical” deviation from the mean. It is a popular measure of variability because it returns to the original units of measure of the data set.
- Like the variance, if the data points are close to the mean, there is a small variation whereas the data points are highly spread out from the mean, then it has a high variance. Standard deviation calculates the extent to which the values differ from the average.
- Standard Deviation, the most widely used measure of dispersion, is based on all values.
- Therefore a change in even one value affects the value of standard deviation. It is independent of origin but not of scale.
- It is also useful in certain advanced statistical problems.

ANOVA

- What is Analysis of Variance (ANOVA)?
- Analysis of variance (ANOVA) is an analysis tool used in statistics that splits an observed aggregate variability found inside a data set into two parts: systematic factors and random factors.
- The systematic factors have a statistical influence on the given data set, while the random factors do not.
- Analysts use the ANOVA test to determine the influence that independent variables have on the dependent variable in a regression study

More details with Standard Deviation and
ANOVA in Next Lecture!!!

- Thanks You !!!

Lecture 12: ANOVA and Standard Deviation

9 Feb 2022

What is Analysis of Variance (ANOVA)?

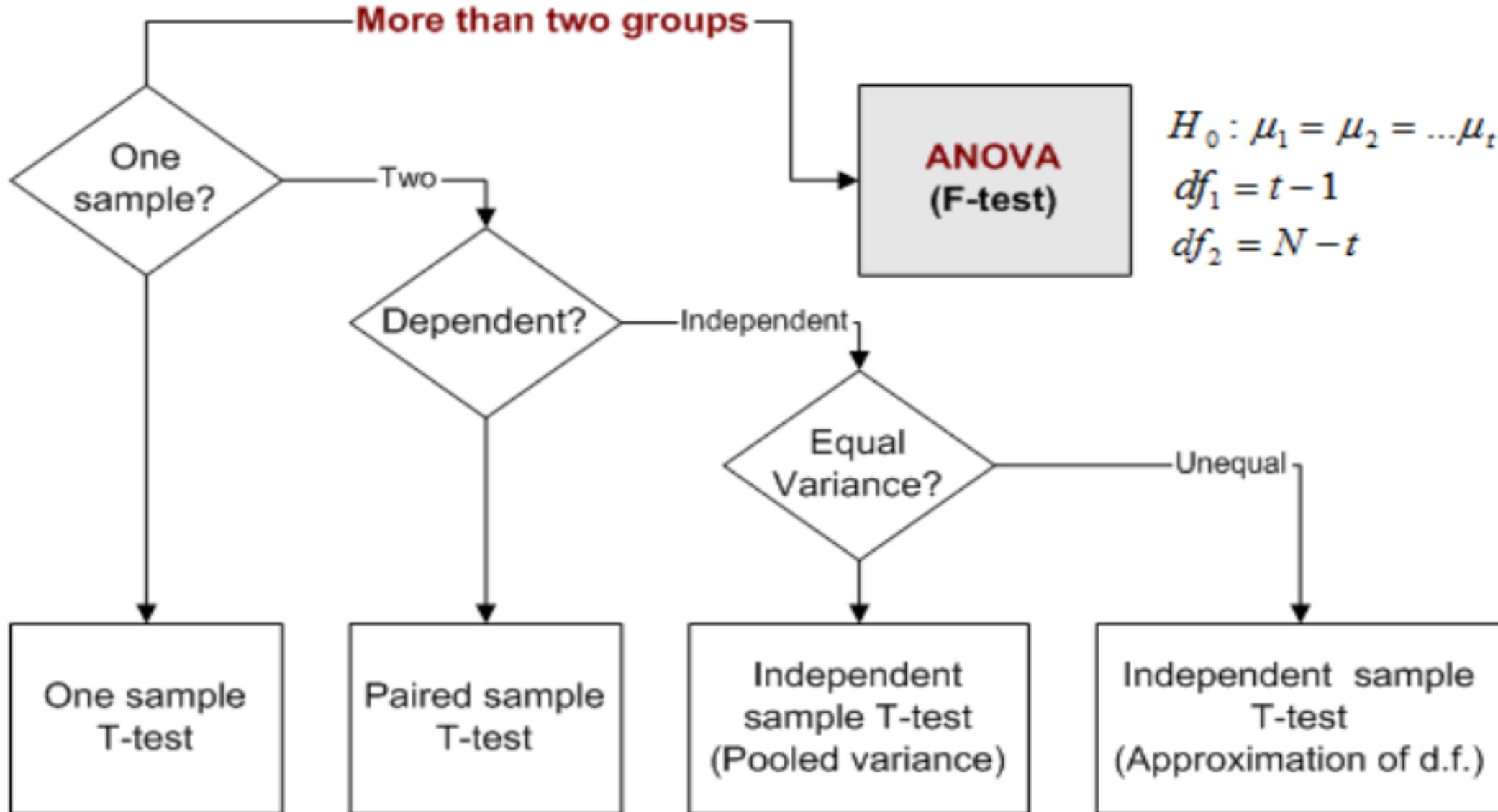
- Analysis of variance (ANOVA) is an analysis tool used in statistics that splits an observed aggregate variability found inside a data set into two parts: systematic factors and random factors.
- The systematic factors have a statistical influence on the given data set, while the random factors do not.
- Analysts use the ANOVA test to determine the influence that independent variables have on the dependent variable in a regression study.

History of ANOVA

- The t- and z-test methods developed in the 20th century were used for statistical analysis until 1918, when Ronald Fisher created the analysis of variance method.
- ANOVA is also called the Fisher analysis of variance, and it is the extension of the t- and z-tests.
- The term became well-known in 1925, after appearing in Fisher's book, "Statistical Methods for Research Workers."
- It was employed in experimental psychology and later expanded to subjects that were more complex.

T Test

- The t-test and ANOVA examine whether group means differ from one another. The t-test compares two groups, while ANOVA can do more than two groups.
- The t-test ANOVA have three assumptions:
 - independence assumption (the elements of one sample are not related to those of the other sample),
 - normality assumption (samples are randomly drawn from the normally distributed populations with unknown population means; otherwise the means are no longer best measures of central tendency, thus test will not be valid), and
 - equal variance assumption (the population variances of the two groups are equal)



$$H_0: \mu = c$$

$$df = n - 1$$

$$H_0: \mu_d = 0$$

$$df = n - 1$$

$$H_0: \mu_1 - \mu_2 = 0$$

$$df = n_1 + n_2 - 2$$

$$H_0: \mu_1 - \mu_2 = 0$$

$$df = \text{approximated}$$

Key Differences Between T-test and ANOVA

- The significant differences between T-test and ANOVA are discussed in detail in the following points:
- T Test -A hypothesis test that is used to compare the means of two populations
- Anova - A statistical technique that is used to compare the means of more than two populations is known as Analysis of Variance or ANOVA.

T Test

- The t-test is any statistical hypothesis test in which the test statistic follows a Student's t-distribution under the null hypothesis.
- It can be used to determine if two sets of data are significantly different from each other, and is most commonly applied when the test statistic would follow a normal distribution if the value of a scaling term in the test statistic were known.
- T-test uses means and standard deviations of two samples to make a comparison

Formula T Test

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\Delta}}$$

where

$$s_{\Delta} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Where,

\bar{x} = Mean of first set of values

\bar{x}_2 = Mean of second set of values

s_1 = Standard deviation of first set of values

s_2 = Standard deviation of second set of values

n_1 = Total number of values in first set

n_2 = Total number of values in second set.

Standard Deviation

The formula for standard deviation is given by:

$$S = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

Where,

x = Values given

\bar{x} = Mean

n = Total number of values.

Example of T Test

- Find the t-test value for the following two sets of values: 7, 2, 9, 8 and 1, 2, 3, 4?

Formula for mean: $\bar{x} = \frac{\sum x}{n}$

Formula for standard deviation: $S = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$

Number of terms in first set: $n_1 = 4$

Mean for first set of data: $\bar{x}_1 = 6.5$

Standard Deviation

x_1	$x_1 - \bar{x}_1$	$(x_1 - \bar{x}_1)^2$
7	0.5	0.25
2	-4.5	20.25
9	2.5	6.25
8	1.5	2.25
		$\sum (x_1 - \bar{x}_1)^2 = 29$

Standard Deviation

- Standard deviation for the first set of data: $S_1 = 3.11$
- Number of terms in second set: $n_2 = 4$
- Mean for second set of data: $x_2 \text{ cap}=2.5$

Standard deviation

x_2	$x_2 - \bar{x}_2$	$(x_2 - \bar{x}_2)^2$
1	-1.5	2.25
2	-0.5	0.25
3	0.5	0.25
4	1.5	2.25
		$\sum (x_2 - \bar{x}_2)^2 = 5$

Standard deviation for first set of data: $s_2S2 = 1.29$

T Test Formula and solution

Formula for t-test value:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\Delta}}$$

where

$$s_{\Delta} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$t = \frac{6.5 - 2.5}{\sqrt{\frac{9.67}{4} + \frac{1.67}{4}}}$$

$$t = 2.3764 = 2.36 \text{ (approx)}$$

Z-Test

- A z-test is a statistical test used to determine whether two population means are different when the variances are known and the sample size is large.
- The test statistic is assumed to have a [normal distribution](#), and nuisance parameters such as standard deviation should be known in order for an accurate z-test to be performed.

Features of Z Test

- A z-test is a statistical test to determine whether two population means are different when the variances are known and the sample size is large.
- A z-test is a hypothesis test in which the z-statistic follows a normal distribution.
- A z-statistic, or z-score, is a number representing the result from the z-test.
- Z-tests are closely related to t-tests, but t-tests are best performed when an experiment has a small sample size.
- Z-tests assume the standard deviation is known, while t-tests assume it is unknown.

Understanding Z-Tests

- The z-test is also a hypothesis test in which the z-statistic follows a normal distribution.
- The z-test is best used for greater-than-30 samples because, under the [central limit theorem](#), as the number of samples gets larger, the samples are considered to be approximately normally distributed.

Central Limit Theorem (CLT)

- In the study of probability theory, the central limit theorem (CLT) states that the distribution of sample approximates a normal distribution (also known as a “bell curve”) as the sample size becomes larger, assuming that all samples are identical in size, and regardless of the population distribution shape.
- Sample sizes equal to or greater than 30 are considered sufficient for the CLT to predict the characteristics of a population accurately.

Z Test

- When conducting a z-test, the null and alternative hypotheses, alpha and z-score should be stated.
- Next, the test statistic should be calculated, and the results and conclusion stated. A z-statistic, or z-score, is a number representing how many standard deviations above or below the mean population a score derived from a z-test is.

Z Score

- A z-score, or z-statistic, is a number representing how many standard deviations above or below the mean population the score derived from a z-test is.
- Essentially, it is a numerical measurement that describes a value's relationship to the mean of a group of values.
- If a z-score is 0, it indicates that the data point's score is identical to the mean score.
- A z-score of 1.0 would indicate a value that is one standard deviation from the mean.
- Z-scores may be positive or negative, with a positive value indicating the score is above the mean and a negative score indicating it is below the mean.

Revision of Z Score

- Z-scores are expressed in terms of standard deviations from their means.
- Resultantly, these z-scores have a distribution with a mean of 0 and a standard deviation of 1.

Formula Z Score

$$\text{Z Score} = (x - \bar{x})/\sigma$$

Where,

- x = Standardized random variable
- \bar{x} = Mean
- σ = Standard deviation.

Example 1:

- **How well did Ram perform in her English coursework compared to the other 50 students?**

	Score (x)	Mean (\bar{x})	Standard Deviation (σ)
English Coursework	70	60	15

Computation of Z Score

$$Z \text{ Score} = (x - \bar{x})/\sigma$$

$$= (70 - 60)/15$$

$$= 10/15$$

$$= 0.6667$$

Example 2: Z score Computation

- A student wrote 2 quizzes. In the first quiz, he scored 80 and in other, he scored 75. The mean and standard deviation of first quiz are 70 and 15 respectively, while the mean and standard deviation of the second quiz are 54 and 12 respectively. The results follow the normal distribution. What can you conclude about the student's result by seeing their z scores?

Step 1: Z score computation for Quiz 1

- Calculation of student's Z score for first quiz:
Standardized random variable $x = 80$
Mean, $\bar{x} = 70$
- Population standard deviation = 15
Formula for Z score is given below:
- $Z \text{ Score} = (x - \bar{x})/\sigma$
- $= (80 - 70) / 15$
- $= 0.667$

Step 2: Z score Computation for second Quiz

- Calculation of student's Z score for second quiz:
- Standardized random variable $x = 75$
- Mean $\bar{x} = 54$
Population standard deviation = 12
- Formula for Z score is given below:
- $Z \text{ Score} = (x - \bar{x})/\sigma$
- $= (75 - 54) / 12$
- $= 1.75$

Step 3: Comparison of Scores

- Since Z score of second quiz is better than that of first quiz, hence it is concluded that he did better in second quiz.

Example of Z Test

- Examples of tests that can be conducted as z-tests include a one-sample location test, a two-sample location test, a paired difference test, and a maximum likelihood estimate.
- Z-tests are closely related to t-tests, but t-tests are best performed when an experiment has a small sample size.
- Also, t-tests assume the standard deviation is unknown, while z-tests assume it is known.
- If the standard deviation of the population is unknown, the assumption of the sample variance equaling the population variance is made.

Key Features of ANOVA

- Analysis of variance, or ANOVA, is a statistical method that separates observed variance data into different components to use for additional tests.
- A one-way ANOVA is used for three or more groups of data, to gain information about the relationship between the dependent and independent variables.
- If no true variance exists between the groups, the ANOVA's F-ratio should equal close to 1.

The Formula for ANOVA is:

$$F = \frac{MST}{MSE}$$

where:

F = ANOVA coefficient

MST = Mean sum of squares due to treatment

MSE = Mean sum of squares due to error

What Does the Analysis of Variance Reveal?

- The ANOVA test is the initial step in analyzing factors that affect a given data set. Once the test is finished, an analyst performs additional testing on the methodical factors that measurably contribute to the data set's inconsistency. The analyst utilizes the ANOVA test results in an f-test to generate additional data that aligns with the proposed [regression](#) models.

ANOVA Test Reveals

- The ANOVA test allows a comparison of more than two groups at the same time to determine whether a relationship exists between them. The result of the ANOVA formula, the F statistic (also called the F-ratio), allows for the analysis of multiple groups of data to determine the variability between samples and within samples.

ANOVA Reveals

- If no real difference exists between the tested groups, which is called the **null hypothesis**, the result of the ANOVA's F-ratio statistic will be close to 1.
- The distribution of all possible values of the F statistic is the F-distribution.
- This is actually a group of distribution functions, with two characteristic numbers, called the numerator **degrees of freedom** and the denominator degrees of freedom.

Example of How to Use ANOVA

- A researcher might, for example, test students from multiple colleges to see if students from one of the colleges consistently outperform students from the other colleges. In a business application, an R&D researcher might test two different processes of creating a product to see if one process is better than the other in terms of cost efficiency.

ANOVA Test

- The type of ANOVA test used depends on a number of factors. It is applied when data needs to be experimental. Analysis of variance is employed if there is no access to statistical software resulting in computing ANOVA by hand. It is simple to use and best suited for small samples. With many experimental designs, the sample sizes have to be the same for the various factor level combinations.
- .

ANOVA Test

- ANOVA is helpful for testing three or more variables.
- It is similar to multiple two-sample t-tests.
- However, it results in fewer type I errors and is appropriate for a range of issues.
- ANOVA groups differences by comparing the means of each group and includes spreading out the variance into diverse sources.
- It is employed with subjects, test groups, between groups and within groups

One-Way ANOVA Versus Two-Way ANOVA

- There are two main types of ANOVA:
 - one-way (or unidirectional)
 - and two-way.
- There are also variations of ANOVA. For example, MANOVA (multivariate ANOVA) differs from ANOVA as the former tests for multiple dependent variables simultaneously while the latter assesses only one dependent variable at a time.
- One-way or two-way refers to the number of independent variables in your analysis of variance test.
- A one-way ANOVA evaluates the impact of a sole factor on a sole response variable. It determines whether all the samples are the same.
- The one-way ANOVA is used to determine whether there are any statistically significant differences between the means of three or more independent (unrelated) groups.

A Two-Way ANOVA

- It is an extension of the one-way ANOVA. With a one-way, you have one independent variable affecting a dependent variable.
- With a two-way ANOVA, there are two independents.
- For example, a two-way ANOVA allows a company to compare worker productivity based on two independent variables, such as salary and skill set.
- It is utilized to observe the interaction between the two factors and tests the effect of two factors at the same time.

ANOVA Revision

- Analysis of variance, or ANOVA, is a strong statistical technique that is used to show the difference between two or more means or components through significance tests.
- It also shows us a way to make multiple comparisons of several populations means.
- The Anova test is performed by comparing two types of variation, the variation between the sample means, as well as the variation within each of the samples.
- :

ANOVA Formula

- The below mentioned formula represents one-way Anova test statistics

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares (MS)	F
Within	$SSW = \sum_{j=1}^k \sum_{i=1}^{l_j} (X_{ij} - \bar{X}_j)^2$	$df_w = k - 1$	$MSW = \frac{SSW}{df_w}$	$F = \frac{MSB}{MSW}$
Between	$SSB = \sum_{j=1}^k (\bar{X}_j - \bar{X})^2$	$df_b = n - k$	$MSB = \frac{SSB}{df_b}$	
Total	$SST = \sum_{j=1}^n (\bar{X}_j - \bar{X})^2$	$df_t = n - 1$		

Terminologies used in Formula

- $F = MST/MSE$
- $MST = SST/ p-1$
- $MSE = SSE/N-p$
- $SSE = \sum (n-1)s^2s^2$
- Where,
- F = Anova Coefficient
- MSB = Mean sum of squares between the groups
- MSW = Mean sum of squares within the groups

MSE = Mean sum of squares due to error

SST = total Sum of squares

p = Total number of populations

n = The total number of samples in a population

SSW = Sum of squares within the groups

SSB = Sum of squares between the groups

SSE = Sum of squares due to error

s = Standard deviation of the samples

N = Total number of observations

EXAMPLE

Types of Animals	Number of animals	Average Domestic animals	Standard Deviation
Dogs	5	12	2
Cats	5	16	1
Hamsters	5	20	4

Calculate the Anova coefficient

Step1:

Construct the following table:

Animal name	n	x	s	s^2
Dogs	5	12	2	4
Cats	5	16	1	1
Hamster	5	20	4	16

Step 2: Computation of SST

- $p = 3$
- $n = 5$
- $N = 15$
- $\bar{x} = 16$
- $SST = \sum n (x - \bar{x})^2$
- $$\begin{aligned} SST &= 5(12-16)^2 + 5(16-16)^2 + 11(20-16)^2 \\ &= 160 \end{aligned}$$

Step3:Computation of MST

$$MST = \frac{SST}{p-1}$$

$$MST = \frac{160}{3-1}$$

$$MST = 80$$

Step 4: Computation of SSE

$$\text{SSE} = \sum (n-1)s^2$$

$$\text{SSE} = 4 \times 4 + 4 \times 1 + 4 \times 16$$

$$\text{SSE} = 84$$

Step 5: Computation of MSE

- $MSE = SSE / (N-p)$

$$= 84 / (15 - 3)$$

$$= 84 / 12$$

$$= 7$$

Final step 6: Computation of Anova coefficient

- $F = MST/MSE$
- $F=80/7$
- $F=11.429$

Additional Question T Test

- The CEO of light bulbs manufacturing company claims that an average light bulb lasts 300 days. A researcher randomly selects 15 bulbs for testing. The sampled bulbs last an average of 290 days, with a standard deviation of 50 days. If the CEO's claim were true, what is the probability that 15 randomly selected bulbs would have an average life of no more than 290 days?

The traditional approach requires you to compute the t statistic, based on data presented in the problem description.

The first thing we need to do is compute the t statistic, based on the following equation:

Where \bar{x} is the sample mean, μ is the population mean, s is the standard deviation of the sample, and n is the sample size.

Using the formula: $t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$ $t = \frac{290 - 300}{\frac{50}{\sqrt{15}}} = \frac{-10}{12.909945} = -0.7745966$

Since we will work with the raw data, we select “Sample mean” from the Random Variable dropdown box.

- The degrees of freedom are equal to $15 - 1 = 14$.
- Assuming the CEO's claim is true, the population mean equals 300.
- The sample mean equals 290.
- The standard deviation of the sample is 50.

The cumulative probability: 0.226. Hence, if the true bulb life were 300 days, there is a 22.6% chance that the average bulb life for 15 randomly selected bulbs would be less than or equal to 290 days

Lecture 13: Week 5_

10/2/2021

What is Skewness?

- Skewness refers to a distortion or asymmetry that deviates from the symmetrical bell curve, or [normal distribution](#), in a set of data.
- If the curve is shifted to the left or to the right, it is said to be skewed. Skewness can be quantified as a representation of the extent to which a given distribution varies from a normal distribution.
- A normal distribution has a skew of zero, while a [lognormal distribution](#), for example, would exhibit some degree of right-skew.

Features of Skewness

- Skewness, in statistics, is the degree of asymmetry observed in a probability distribution.
- Distributions can exhibit right (positive) skewness or left (negative) skewness to varying degrees. A normal distribution (bell curve) exhibits zero skewness.
- Investors note right-skewness when judging a return distribution because it, like excess kurtosis, better represents the extremes of the data set rather than focusing solely on the average.

Details on Skewness

- Besides positive and negative skew, distributions can also be said to have zero or undefined skew.
- In the curve of a distribution, the data on the right side of the curve may taper differently from the data on the left side.
- These tapering's are known as "tails."
- Negative skew refers to a longer or fatter tail on the left side of the distribution, while positive skew refers to a longer or fatter tail on the right.

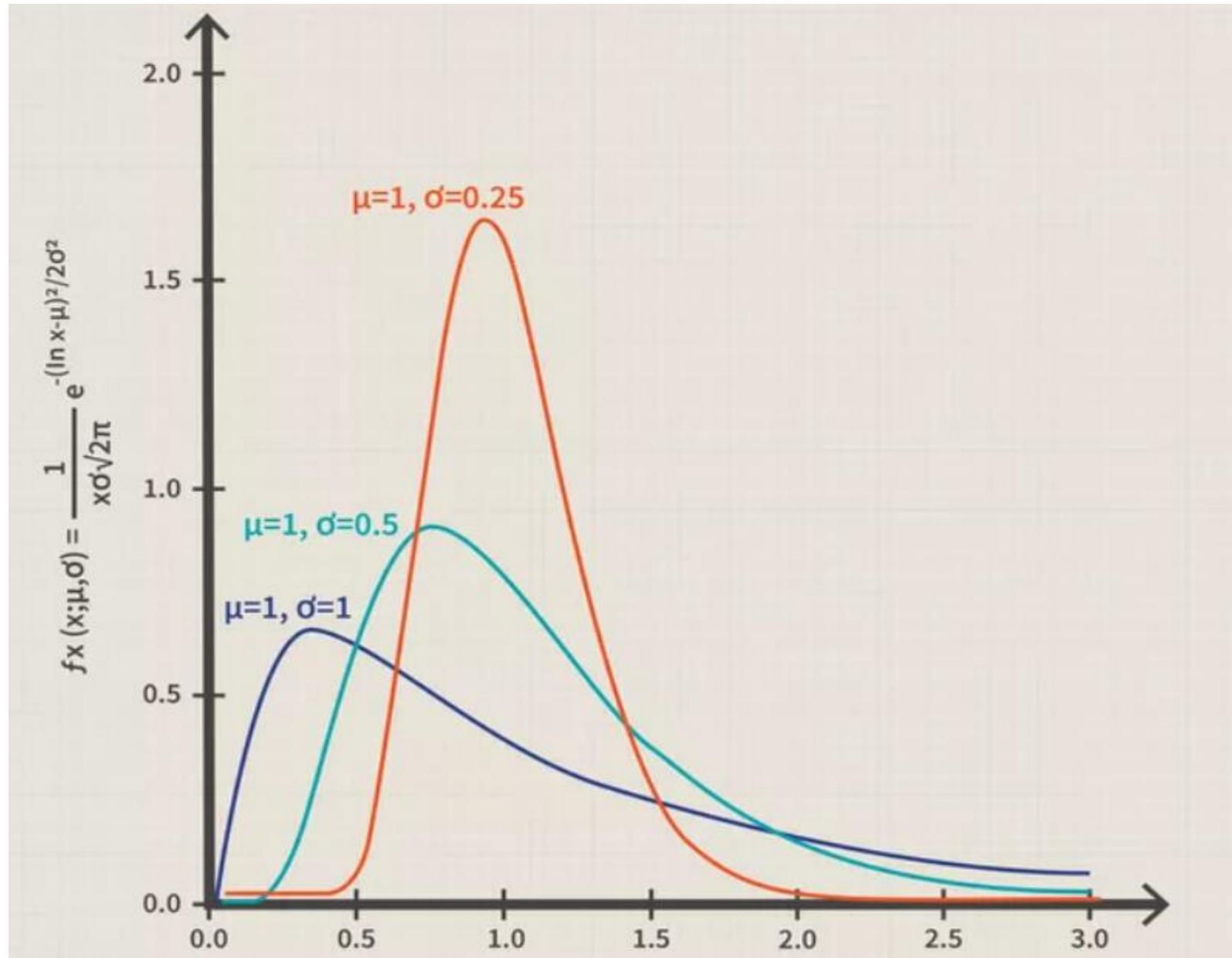
Details on Skewness

- The mean of positively skewed data will be greater than the median.
- In a distribution that is negatively skewed, the exact opposite is the case: the mean of negatively skewed data will be less than the median.
- If the data graphs symmetrically, the distribution has zero skewness, regardless of how long or fat the tails are.

Details on Skewness

- The three probability distributions depicted below are positively-skewed (or right-skewed) to an increasing degree.
- Negatively-skewed distributions are also known as left-skewed distributions.

Skewness



Measures of Skewness:

- Skewness formula is called so because the graph plotted is displayed in skewed manner.
- Skewness is a measure used in statistics that helps reveal the asymmetry of a probability distribution.
- It can either be positive or negative, irrespective of signs.
- To calculate the skewness, we have to first find the mean and variance of the given data.

Calculate Skewness

- Step 1: Find the mean and variance of the given data
- Step 2: The formula for Sample Skewness is :

$$g = \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{(n-1)s^3}$$

Where,

\bar{x} is the sample mean

x_i is the ith sample

n is the total number of observations

s is the standard deviation

g = sample skewness

Find the skewness in the following data

Height (inches)	Class Marks	Frequency
59.5 – 62.5	61	5
62.5 – 65.5	64	18
65.5 – 68.5	67	42
68.5 – 71.5	70	27
71.5 – 74.5	73	8

Step 1:

- To know how skewed these data are as compared to other data sets, we have to compute the skewness.
- Sample size and sample mean should be found out.
- $N = 5 + 18 + 42 + 27 + 8 = 100$

$$\bar{x} = \frac{(61 \times 5) + (64 \times 18) + (67 \times 42) + (70 \times 27) + (73 \times 8)}{100}$$

$$\bar{x} = \frac{6745}{100} = 67.45$$

Step 2: Compute the skewness

Class Mark, x	Frequency, f	xf	$(x - \bar{x})$	$(x - \bar{x})^2 \times f$	$(x - \bar{x})^3 \times f$
61	5	305	-6.45	208.01	-1341.68
64	18	1152	-3.45	214.25	-739.15
67	42	2814	-0.45	8.51	-3.83
70	27	1890	2.55	175.57	447.70
73	8	584	5.55	246.42	1367.63
		6745	n/a	852.75	-269.33
		67.45	n/a	8.5275	-2.6933

Now, the skewness is

$$g = \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{(n-1)s^3}$$

$$s = \sqrt{[(8.5275/(100-1))]} = 0.2935$$

$$g = \sqrt{[-2.693/[99 * (0.295)^3]]} = -1.038$$

For interpreting we have the following rules as per Bulmer in the year 1979:

- If the skewness comes to less than -1 or greater than +1, the data distribution is highly skewed
- If the skewness comes to between -1 and $-\frac{1}{2}$ or between $+\frac{1}{2}$ and +1, the data distribution is moderately skewed.
- If the skewness is between $-\frac{1}{2}$ and $+\frac{1}{2}$, the distribution is approximately symmetric

Another Method to measure Skewness

- Pearson's first and second coefficients of skewness are two common ones.
- Pearson's first coefficient of skewness, or Pearson mode skewness, subtracts the mode from the mean and divides the difference by the standard deviation.
- Pearson's second coefficient of skewness, or Pearson median skewness, subtracts the median from the mean, multiplies the difference by three, and divides the product by the standard deviation.

The Formulae for Pearson's Skewness Are:

$$Sk_1 = \frac{X - Mo}{s}$$

$$Sk_2 = \frac{3\bar{X} - Md}{s}$$

where:

Sk_1 = Pearson's first coefficient of skewness and Sk_2 the second

s = the standard deviation for the sample

\bar{X} = is the mean value

Mo = the modal (mode) value

Md = is the median value

Importance of Pearson's Coefficient

- Pearson's first coefficient of skewness is useful if the data exhibit a strong mode.
- If the data have a weak mode or multiple modes, Pearson's second coefficient may be preferable, as it does not rely on mode as a measure of central tendency.

Practical Importance of Skewness

- Let it be the example of Investors .
- Investor note skewness when judging a return distribution because it, like kurtosis, considers the extremes of the data set rather than focusing solely on the average.
- Short- and medium-term investors in particular need to look at extremes because they are less likely to hold a position long enough to be confident that the average will work itself out.
- Investors commonly use standard deviation to predict future returns, but the standard deviation assumes a normal distribution.

Risk with Skewness

- As few return distributions come close to normal, skewness is a better measure on which to base performance predictions. This is due to skewness risk.
- Skewness risk is the increased risk of turning up a data point of high skewness in a skewed distribution.
- Many financial models that attempt to predict the future performance of an [asset](#) assume a normal distribution, in which measures of central tendency are equal.
- If the data are skewed, this kind of model will always underestimate skewness risk in its predictions.
- The more skewed the data, the less accurate this financial model will be.

Bowley's Coefficient of Skewness for Ungrouped data

- Skewness is a measure of symmetry. The meaning of skewness is “lack of symmetry”.
- Skewness gives us an idea about the concentration of higher or lower data values around the central value of the data.

Bowley's Coefficient

- Bowley's Coefficient of Skewness is also known as Quartile Coefficient of skewness.
- Specially used where quartile and medians are used
 - When the mode is ill –defined and extreme observations are present in the data
 - When distributions has open end classes and unequal class intervals
- In these situations Pearson coefficient of skewness cannot be used.

Bowley's Coefficient

For a symmetric distribution, the two quartiles namely Q_1 and Q_3 are equidistance from the median (i.e. Q_2). That is for symmetric distribution $Q_3 - Q_2 = Q_2 - Q_1$.

If the distribution is not symmetric (i.e., skewed) then the distance $Q_3 - Q_2$ is not equal to the distance $Q_2 - Q_1$. That is for asymmetric distribution $Q_3 - Q_2 \neq Q_2 - Q_1$.

The absolute measure of skewness is $(Q_3 - Q_2) - (Q_2 - Q_1) = Q_3 + Q_1 - 2 * Q_2$.

Formula for Bowley's Coefficient

Bowley's coefficient of skewness is the relative measure of skewness. It is denoted by S_b and is defined as

$$S_b = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$

where,

- Q_1 is the first quartile,
- Q_2 is the second quartile,
- Q_3 is the third quartile,

Types of Skewness

- If $S_b < 0$, i.e., $Q_3 - Q_2 < Q_2 - Q_1$ then the distribution is **negatively skewed**.
- If $S_b = 0$, i.e., $Q_3 - Q_2 = Q_2 - Q_1$ then the distribution is **Symmetric** or **not skewed**.
- If $S_b > 0$, i.e., $Q_3 - Q_2 > Q_2 - Q_1$ then the distribution is **positively skewed**.

Bowley's coefficient of skewness ranges from -1 to +1

Example

A random sample of 15 patients yielded the following data on the length of stay (in days) in the hospital.

5, 6, 9, 10, 15, 10, 14, 12, 10, 13, 13, 9, 8, 10, 12.

Find Bowley's coefficient of skewness.

Step 1:

The formula for i^{th} quartile is

$$Q_i = \text{Value of } \left(\frac{i(n+1)}{4} \right)^{\text{th}}$$

observation, $i=1,2,3$

where n is the total number of observations.

Arrange the data in ascending order

5, 6, 8, 9, 9, 10, 10, 10, 12, 12, 13, 13, 14, 15

Step 2: First Quartile Q₁

The first quartile Q_1 can be computed as follows:

$$\begin{aligned} Q_1 &= \text{Value of } \left(\frac{1(n+1)}{4} \right)^{\text{th}} \text{ obs.} \\ &= \text{Value of } \left(\frac{1(15+1)}{4} \right)^{\text{th}} \text{ obs.} \\ &= \text{Value of } (4)^{\text{th}} \text{ obs.} \\ &= 9 \text{ days.} \end{aligned}$$

Thus, lower 25 % of the patients had length of stay in the hospital less than or equal to 9 days.

Second Quartile

The second quartile Q_2 can be computed as follows:

$$\begin{aligned} Q_2 &= \text{Value of } \left(\frac{2(n+1)}{4} \right)^{\text{th}} \text{ obs.} \\ &= \text{Value of } \left(\frac{2(15+1)}{4} \right)^{\text{th}} \text{ obs.} \\ &= \text{Value of } (8)^{\text{th}} \text{ obs.} \\ &= 10 \text{ days.} \end{aligned}$$

Thus, lower 50 % of the patients had length of stay in the hospital less than or equal to 10 days.

Third Quartile

The third quartile Q_3 can be computed as follows:

$$\begin{aligned} Q_3 &= \text{Value of } \left(\frac{3(n+1)}{4} \right)^{\text{th}} \text{ obs.} \\ &= \text{Value of } \left(\frac{3(15+1)}{4} \right)^{\text{th}} \text{ obs.} \\ &= \text{Value of } (12)^{\text{th}} \text{ obs.} \\ &= 13 \text{ days.} \end{aligned}$$

Thus, lower 75 % of the patients had length of stay in the hospital less than or equal to 13 days.

Step 4: Computation of Bowley's coefficient

Bowley's coefficient of skewness

Bowley's coefficient of skewness is

$$\begin{aligned} S_b &= \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1} \\ &= \frac{13 + 9 - 2 * 10}{13 - 9} \\ &= 0.5 \end{aligned}$$

As the coefficient of skewness $S_b > 0$, the data is positively skewed.

Measure of Distance

Bregman divergence

- In mathematics, specifically statistics and information geometry, a **Bregman divergence** or **Bregman distance** is a measure of difference between two points, defined in terms of a strictly convex function; they form an important class of divergences. When the points are interpreted as probability distributions – notably as either values of the parameter of a parametric model or as a data set of observed values – the resulting distance is a statistical distance . The most basic Bregman divergence is the squared Euclidean distance.

History of Bregman divergences

- Bregman divergences are similar to [metrics](#), but satisfy neither the [triangle inequality](#) (ever) nor symmetry (in general). However, they satisfy a generalization of the [Pythagorean theorem](#), and in information geometry the corresponding [statistical manifold](#) is interpreted as a (dually) flat [manifold](#). This allows many techniques of [optimization theory](#) to be generalized to Bregman divergences, geometrically as generalizations of [least squares](#).
- Bregman divergences are named after [Lev M. Bregman](#), who introduced the concept in 1967.

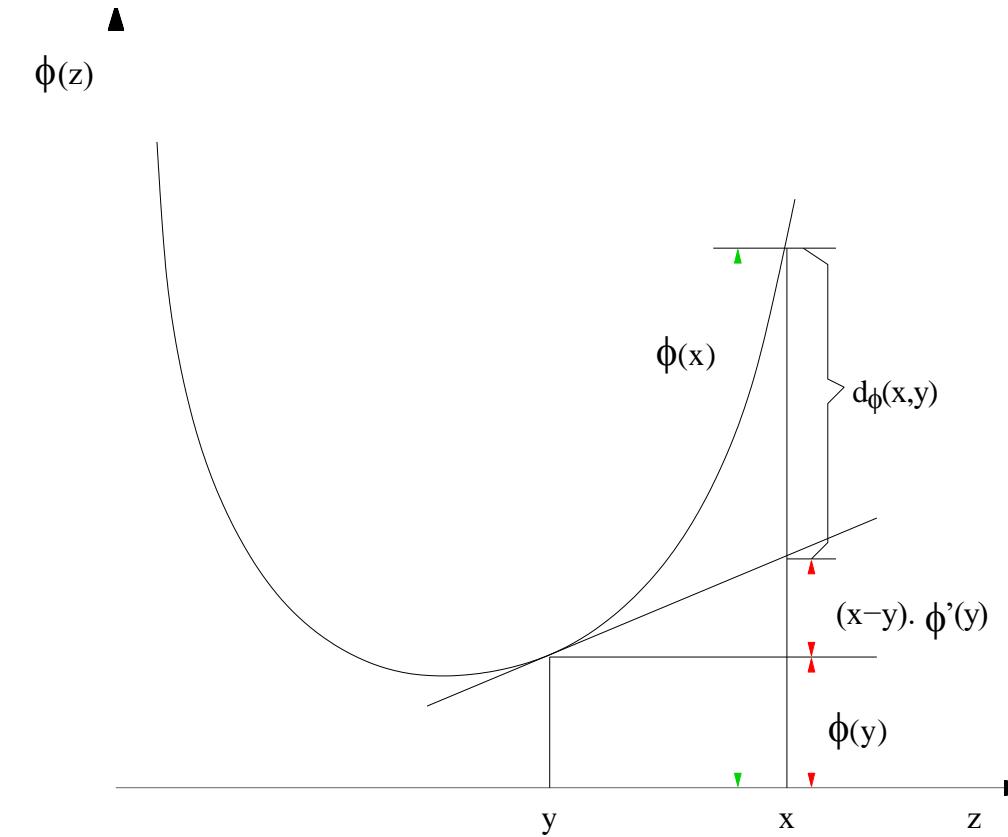
Applications of Bregman divergences

- In machine learning, Bregman divergences are used to calculate the bi-tempered logistic loss, performing better than the **softmax function** with noisy datasets

Bregman divergence

- Let $\psi : \Omega \rightarrow \mathbb{R}$ be a function that is:
 - a) strictly convex,
 - b) continuously differentiable,
 - c) defined on a closed convex set Ω .
- Then the Bregman divergence is defined as
$$\Delta\psi(x, y) = (\psi(x) - \psi(y)) - (\nabla\psi(y), x - y), \forall x, y \in \Omega.$$
- That is, the difference between the value of ψ at x and the first order Taylor expansion of ψ around y evaluated at point x .

Bregman Divergences



φ is strictly convex, differentiable

$$d_\varphi(x, y) = \varphi(x) - \varphi(y) - (x - y, \nabla\varphi(y))$$

-p.417/?

Examples

- $\varphi(x) = \|x\|^2$ is strictly convex and differentiable on \mathbb{R}^m
 - $d_\varphi(x, y) = \|x - y\|^2$ [squared Euclidean distance]
- $\varphi(p) = \sum_{j=1}^m p_j \log p_j$ (negative entropy) is strictly convex and differentiable on the m -simplex
 - $d_\varphi(p, q) = \sum_{j=1}^m p_j \log \frac{p_j}{q_j}$ [KL-divergence]
- $\varphi(x) = -\sum_{j=1}^m \log x_j$ is strictly convex and differentiable on \mathbb{R}_{++}^m
 - $d_\varphi(x, y) = \sum_{j=1}^m \left| \frac{x_j}{y_j} - \log \frac{x_j}{y_j} + 1 \right|$ [Itakura-Saito distance]

Properties of Bregman Divergences

$d_\varphi(x, y) \geq 0$, and equals 0 iff $x = y$, but not a metric (symmetry,



triangle inequality do not hold)



Convex in the first argument, but not necessarily in the second one



KL divergence between two distributions of the same exponential family is a Bregman divergence



Generalized Law of Cosines and Pythagoras Theorem:

$$d_\varphi(x, y) = d_\varphi(z, y) + d_\varphi(x, z) - ((x - z), (\nabla\varphi(y) - \nabla\varphi(z)))$$

When $x \in$ convex (affine) set Ω & z is the Bregman projection onto Ω

$$z \equiv P_\Omega(y) = \operatorname{argmin}_{\omega \in \Omega} d_\varphi(\omega, y),$$

the inner product term becomes negative (equals zero)

Bregman Information

- For squared loss
 - Mean is the best constant predictor of a random variable

$$\mu = \operatorname{argmin}_{\mathbf{c}} E[\|X - \mathbf{c}\|^2]$$

- The minimum loss is the variance $E[\|X - \mu\|^2]$
- Theorem: For all Bregman divergences

$$\mu = \operatorname{argmin}_{\mathbf{c}} E[d_\varphi(X, \mathbf{c})]$$

- Definition: The minimum loss is the Bregman information of X

$$I_\varphi(X) = E[d_\varphi(X, \mu)]$$

- (minimum distortion at Rate = 0)

Examples of Bregman Information

- $\varphi(x) = \|x\|^2, X \sim v$ over \mathbb{R}^m
 - $I_\varphi(X) = E_v [\|X - E_v [X]\|^2]$ [Variance]
- $\varphi(x) = \sum_{j=1}^m x_j \log x_j, X \sim p(z)$ over $\{p(Y|z)\} \subset m\text{-simplex}$
 - $I_\varphi(X) = I(Z; Y)$ [Mutual Information]
- $\varphi(x) = -\sum_{j=1}^m \log x_j, X \sim \text{uniform over } \{x^{(i)}\}_{i=1}^n \subset \mathbb{R}^m$
 - $I_\varphi(X) = \sum_{j=1}^m \log \frac{\mu_j}{g_j}$ [log AM/GM]

Bregman Divergence

$$D_\varphi(\mathbf{x}, \mathbf{y}) = \varphi(\mathbf{x}) - \varphi(\mathbf{y}) - (\mathbf{x} - \mathbf{y})^T \nabla \varphi(\mathbf{y})$$

- Bregman distance is a measure of distance between two points, defined in terms of a strictly convex function. When the points are interpreted as probability distributions – notably as either values of the parameter of a parametric model or as a data set of observed values – the resulting distance is calculated using Bregman divergence.
- Generalize squared Euclidean distance to a class of distances that all share similar properties
- It's a family of proximity functions that have common properties. It represents loss or distortion functions.
- What is a loss function?
- Let \mathbf{x} and \mathbf{y} be two points, where \mathbf{y} is regarded as the original point and \mathbf{x} is some distortion or approximation of it. \mathbf{x} may be a point that was generated by adding random noise to \mathbf{y} . The goal is to measure the resulting distortion or loss that results if \mathbf{y} is approximated by \mathbf{x} . Of course, the more similar \mathbf{x} and \mathbf{y} are, the smaller the loss or distortion.
- Bregman divergences can be used as dissimilarity functions. Bregman divergence (loss function) $D(\mathbf{x}, \mathbf{y})$ generated by that function is given by the following equation: $D(\mathbf{x}, \mathbf{y}) = \Phi(\mathbf{x}) - \Phi(\mathbf{y}) - \langle \nabla \Phi(\mathbf{y}), (\mathbf{x} - \mathbf{y}) \rangle$
 - Φ - a strictly convex function
 - $\nabla \Phi(\mathbf{y})$ is the gradient of Φ evaluated at \mathbf{y}
 - $\mathbf{x}-\mathbf{y}$ is the vector difference between \mathbf{x} and \mathbf{y}
 - $\langle \nabla \Phi(\mathbf{y}), (\mathbf{x} - \mathbf{y}) \rangle$ is the inner product between $\nabla \Phi(\mathbf{y})$ and $(\mathbf{x} - \mathbf{y})$. For points in Euclidean space, the inner product is just the dot product.
- $d^2(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^2 = \langle \mathbf{x} - \mathbf{y}, \mathbf{x} - \mathbf{y} \rangle = \|\mathbf{x}\|^2 - \|\mathbf{y}\|^2 - 2\langle \mathbf{y}, \mathbf{x} - \mathbf{y} \rangle$. Derived using the euclidean distance formula.

Bregman Divergence

- Let x and y be real numbers and $\phi(t)$ be the real valued function, $\phi(t) = t^2$. The gradient reduces to the derivative and the dot product reduces to multiplication. $D(x, y) = x^2 - y^2 - 2y(x - y) = (x - y)^2$
- The graph for this example, with $y = 1$, is shown for two values of x : $x = 2$ and $x = 3$.

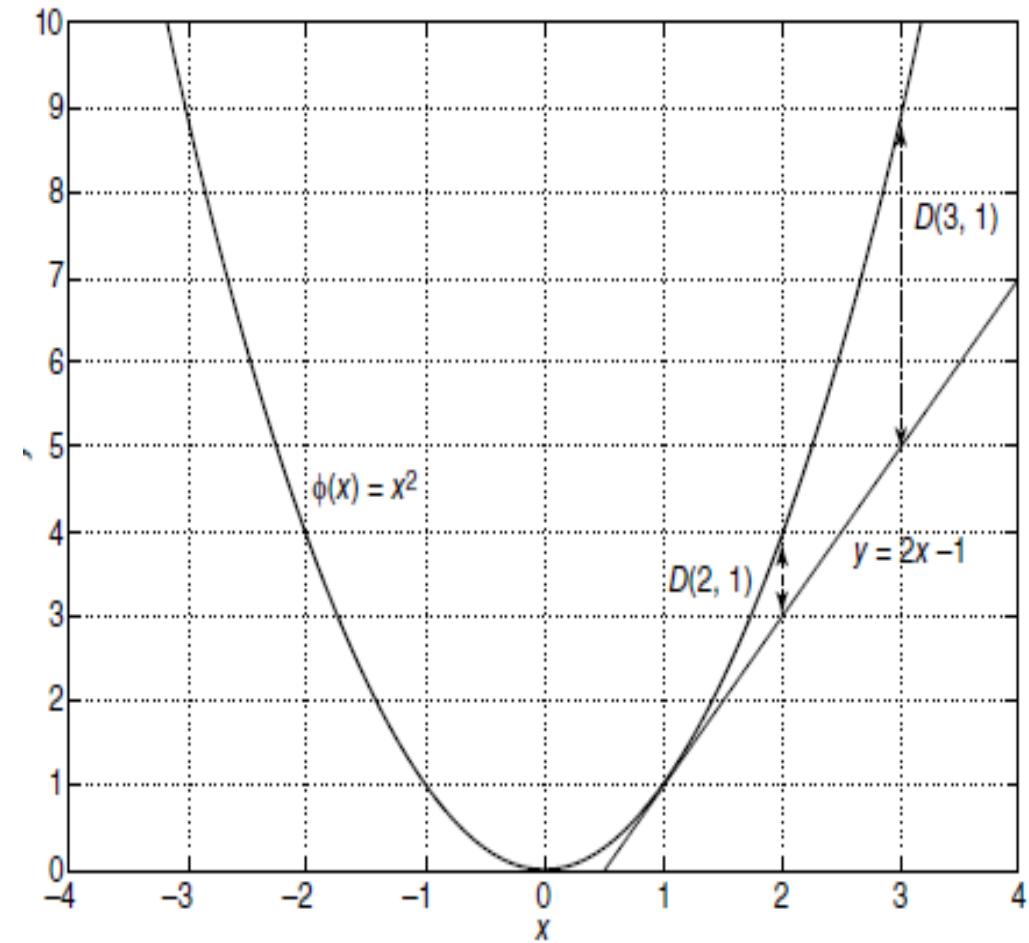
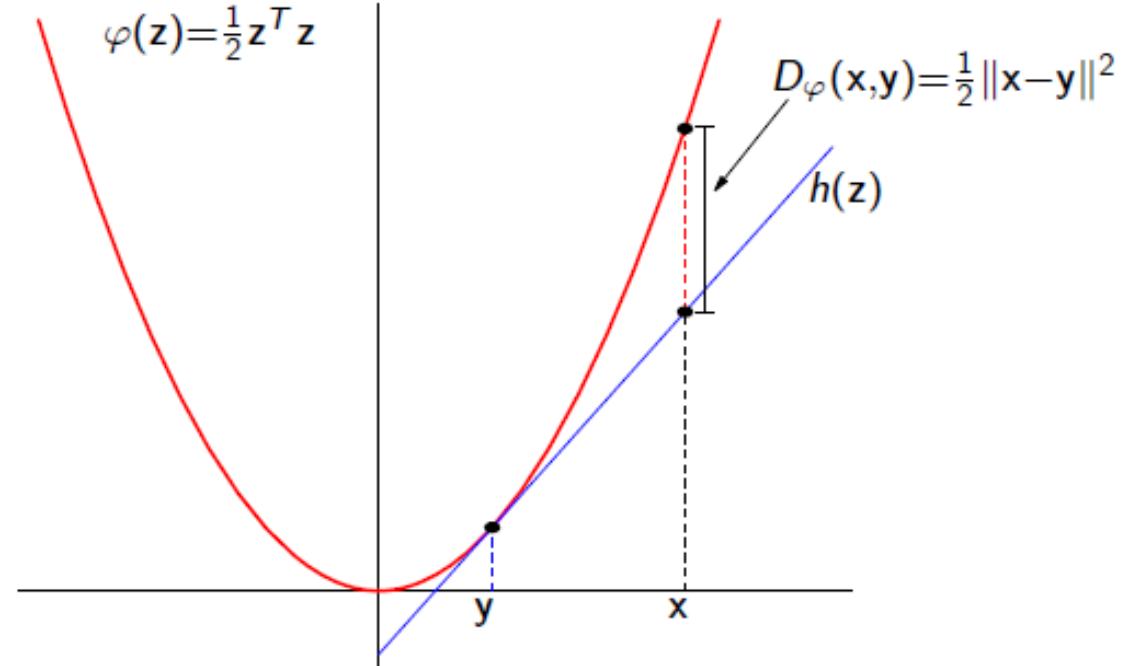


Figure 2.18. Illustration of Bregman divergence.

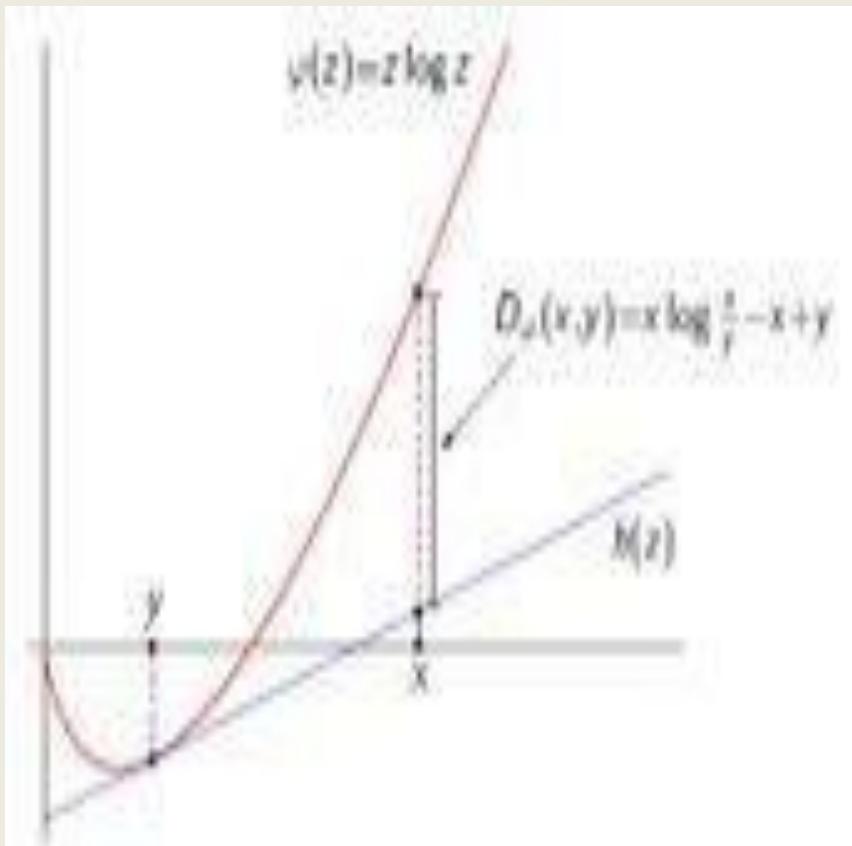
Types of Bregman Divergence

1. Squared Euclidean distance is a Bregman divergence



Kullback-Leibler (KL) divergence

- Relative Entropy (or KL-divergence) is another Bregman divergence using the convex function $f_{KL}(p) = \sum_{i=1}^n p_i \log p_i$



$$D_\varphi(x,y) = x \log \frac{x}{y} - x + y$$

$$D_\varphi(\mathbf{x},\mathbf{y}) = \frac{1}{2} \|\mathbf{x} - \mathbf{y}\|^2$$

$$D_\varphi(x,y) = \frac{x}{y} - \log \frac{x}{y} - 1$$

Properties Of Bregma n Divergence

- **Non-negativity:** $D_F(p, q) \geq 0$ for all p, q . This is a consequence of the convexity of F
- **Convexity:** $D_F(p, q)$ is convex in its first argument, but not necessarily in the second argument
- **Linearity**
- **Mean as minimizer:** A key result about Bregman divergences is that, given a random vector, the mean vector minimizes the expected Bregman divergence from the random vector

Topic Covered : Measures of Skewness: Pearson's coefficient, Bowley's coefficient, coefficient based upon moments

-

Lecture 14: Week 6_

14/2/2021

Agenda

- Coefficient based upon moments
- Similarity Measure Vs Dissimilarity Measure
- Minkowski distance
- Euclidean distance
- manhattan distance
- Supremum distance
- Mahalanobis distance
- Bhattacharyya distance

Coefficient Based Upon Moments

- The coefficient of skewness measures the skewness of a distribution.
- It is based on the notion of the **moment** of the distribution. This coefficient is one of the **measures of skewness**.
- Central Moments- The average of all the deviations of all observations in a dataset from the mean of the observations raised to the power r

Mean Moments

In mean moments, the deviations are taken from the mean.

For Ungrouped Data:

$$\text{First Population Moment about Mean} = \mu_1 = \frac{\sum(x_i - \mu)}{N}$$

$$\text{Second Population Moment about Mean} = \mu_2 = \frac{\sum(x_i - \mu)^2}{N}$$

$$\text{First Sample Moment about Mean} = m_1 = \frac{\sum(x_i - \bar{x})}{n}$$

$$\text{Second Sample Moment about Mean} = m_2 = \frac{\sum(x_i - \bar{x})^2}{n}$$

In General,

$$r^{th} \text{ Population Moment about Mean} = \mu_r = \frac{\sum(x_i - \mu)^r}{N}$$

$$r^{th} \text{ Sample Moment about Mean} = m_r = \frac{\sum(x_i - \bar{x})^r}{n}$$

Central(Mean)Moments

- **Formula for Grouped Data:**

$$r^{th} \text{ Population Moment about Mean} = \mu_r = \frac{\sum f(x_i - \mu)^r}{\sum f}$$

$$r^{th} \text{ Sample Moment about Mean} = m_r = \frac{\sum f(x_i - \bar{x})^r}{\sum f}$$

Types of Central Moments

There are 4 central moments:

- The first central moment, $r=1$, is the sum of the difference of each observation from the sample average (arithmetic mean), which always equals 0
- The second central moment, $r=2$, is variance.

Types of Central Moments

The third central moment, $r=3$, is skewness.

Skewness describes how the sample differs in shape from a symmetrical distribution.

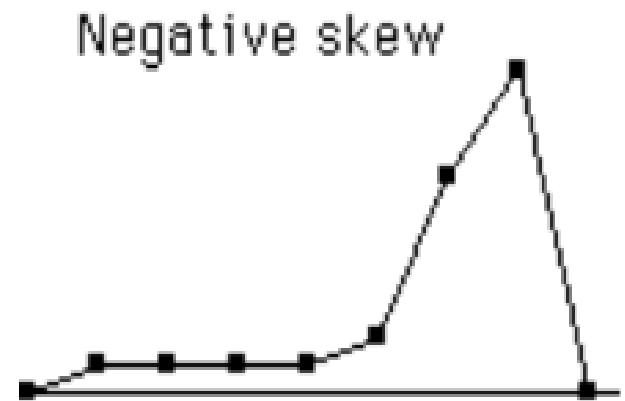
If a normal distribution has a skewness of 0, right skewed is greater than 0 and left skewed is less than 0.

Negative Skewed Distribution -Revision

Negatively skewed distributions, skewed to the left, occur when most of the scores are towards the left of the mode of the distribution.

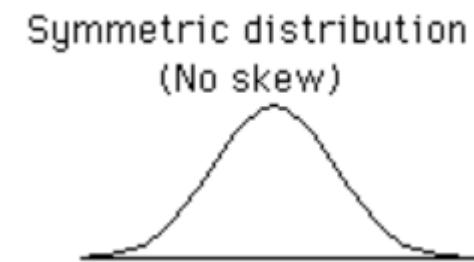
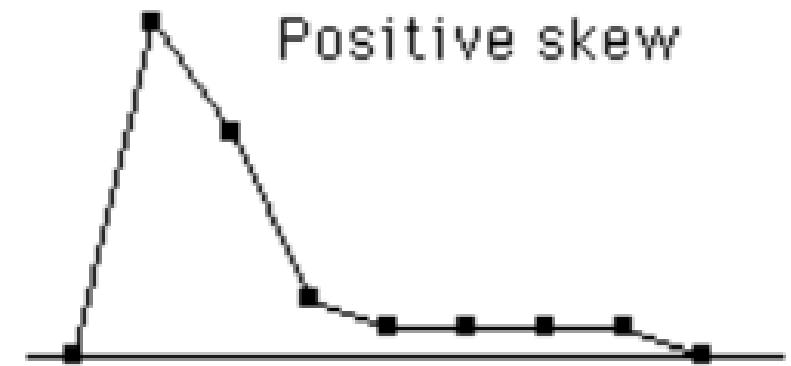
In a normal distribution where skewness is 0, the mean, median and mode are equal.

In a negatively skewed distribution, the mode > median > mean.



Positively skewed distributions

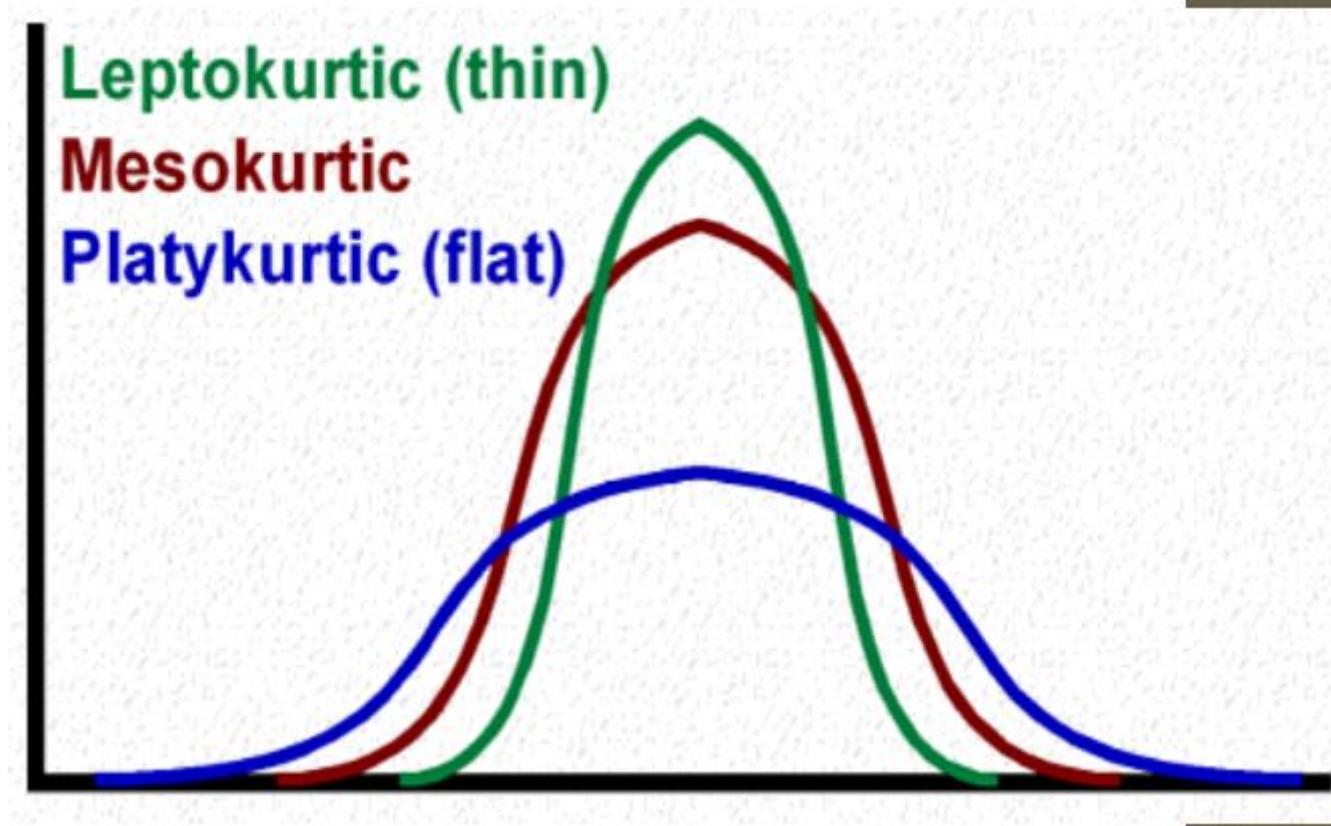
- Positively skewed distributions occur when most of the scores are towards the right of the mode of the distribution. In a positively skewed distribution, $\text{mode} < \text{median} < \text{mean}$



Kurtosis

- Karl Pearson introduced the term Kurtosis (literally the amount of hump) for the degree of peakedness or flatness of a unimodal frequency curve. When the peak of a curve becomes relatively high then that curve is called Leptokurtic. When the curve is flat-topped, then it is called Platykurtic. Since normal curve is neither very peaked nor very flat topped, so it is taken as a basis for comparison. The normal curve is called Mesokurtic

Kurtosis-Diagram



Kurtosis central moment

- Kurtosis is the 4th central moment.
- This is the “peakedness” of a distribution. It measures the extent to which the data are distributed in the tails versus the center of the distribution
- There are three types of peakedness. Leptokurtic- very peaked
Platykurtic – relatively flat Mesokurtic – in between

Example:

- The Frequency distributions of scores in Exploratory Data Analysis subject for 50 students are as follows:

Score	50-60	60-70	70-80	80-90	90-100	100-110	110-120	120-130	130-140
Frequency	1	0	0	1	1	2	1	0	4
Score	140-150	150-160	160-170	170-180	180-190	190-200	200-210	210-220	220-230
Frequency	4	2	5	10	11	4	1	1	2

- Use Above data and compute the First Four Moment about Mean and then find moment of coefficient of skewness and Kurtosis and Comments on nature of distributions?

Step 1

Mid Value (x)	f	d= (x-135)/10	f*d	f*d*d	f*d*d*d	f*d*d*d*d
55	1	-8	-8	64	-512	4096
65	0	-7	0	0	0	0
75	0	-6	0	0	0	0
85	1	-5	-5	25	-125	625
95	1	-4	-4	16	-64	256
105	2	-3	-6	18	-54	162
115	1	-2	-2	4	-8	16
125	0	-1	0	0	0	0
135	4	0	0	0	0	0
145	4	1	4	4	4	4
155	2	2	4	8	16	32
165	5	3	15	45	135	405
175	10	4	40	160	640	2560
185	11	5	55	275	1375	6875
195	4	6	24	144	864	5184
205	1	7	7	49	343	2401
215	1	8	8	64	512	4096
225	2	9	18	162	1458	13122
Total	50		150	1038	4584	39834

Step 2: Computation of the raw moment

The raw moment	$M_1 = 150/50$	
$M_1 = \sum fd / \sum f$		3
$M_2 = \sum fd * d / \sum f$	$M_2 = 1038/50$	
$M_3 = \sum fd * d * d / \sum f$	$M_3 = 4584/50$	20.76
$M_3 = \sum fd * d * d * d / \sum f$	$M_4 = 39834/50$	91.68
		796.68

Step 3: Computation of Central Moment

height=h=10		
Central Moments of variable x $(M2-M1*M1)*h*h$	m2=	1176
Central Moments of variable x $(M3-3M2*M1)+ 2M1*M1*M1* h*h*h$	m3=	-41160
Central Moments of variable x $(M4-4M3*M1+6 M2*M1*M1-3 M1*M1*M1*M1)* h*h*h*h$	m4=	5745600

Step 4:Compute Moment Coefficient of Skewness= Y1 and comment

$$Y_1 = M_3 / (M_2 + \sqrt{M_2})$$

$$\begin{aligned} & (\text{minus}) 41160 / 1176 * \\ & \quad \text{square root of } (1176) \\ & \quad = -1.02 \end{aligned}$$

It means that Negative skewed

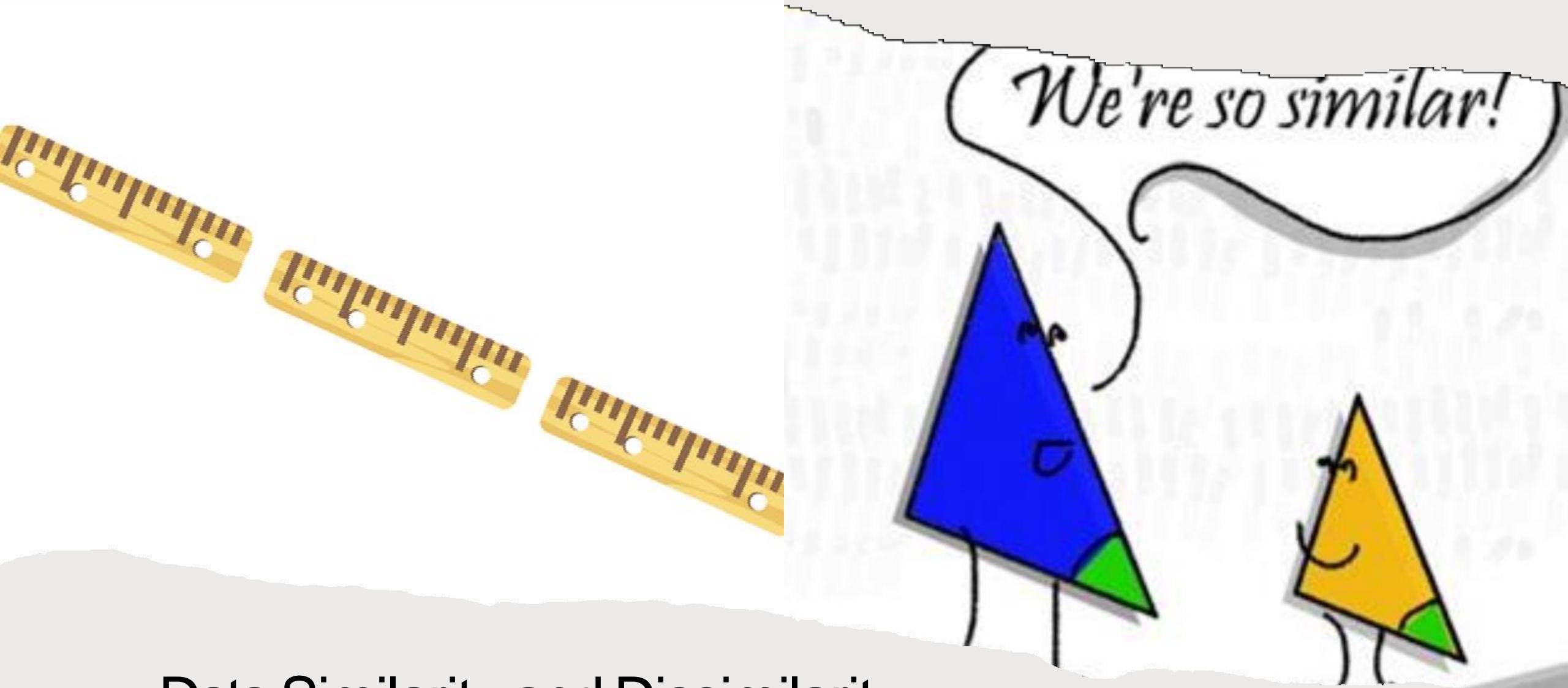
The frequency curve of the given distribution has longer tail toward the left

Moment of Coefficient of Kurtosis (Beta2)

- Beta 2=M4/M2*M2=5745600/(1176)*1176=4.15
- Interpretation:
- Beta 2 is greater than 3 (More than M1) . Hence the distribution is leptokurtic .
- That means Frequency Curve is more peaked (thin) than the normal curve.

Lecture 15: Week 6_

15/2/2021



We're so similar!

Data Similarity and Dissimilarity

What is Similarity?

The quality or state of being similar; likeness; resemblance; as, a similarity of features.

Webster's Dictionary



Similarity is hard to define, but...
“We know it when we see it”

The real meaning of similarity is a philosophical question. We will take a more pragmatic approach.

Similarity and Dissimilarity

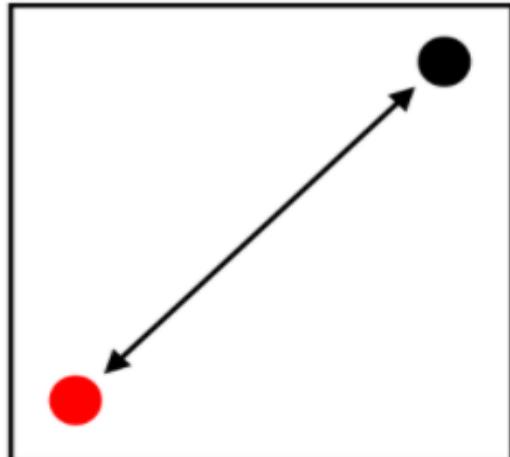
- **Similarity is the quantity that reflects strength of relationship between 2 objects or features.**
- **Similarity is difficult to measure.**
- Dissimilarity measures the discrepancy between 2 objects based on several features. It is a measure of dissimilarity.
- Distance measures dissimilarity.
- When similarity is 1 dissimilarity is 0 and similarity is 0 dissimilarity is 1.
- $\text{Sim} = 1 - \text{Disim}$

Similarity Measure Vs Dissimilarity Measure

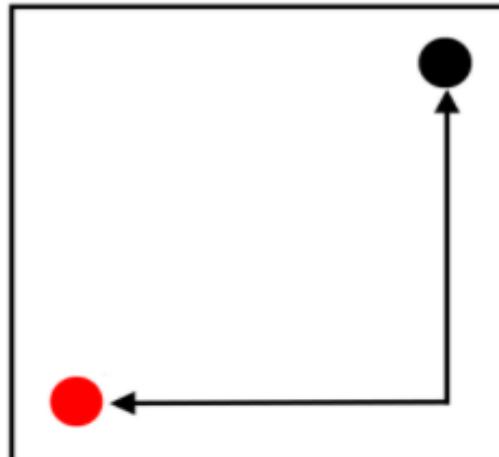
- In data science, the similarity measure is a way of measuring how data samples are related or closed to each other.
- On the other hand, the dissimilarity measure is to tell how much the data objects are distinct.
- Moreover, these terms are often used in clustering when similar data samples are grouped into one cluster.
- All other data samples are grouped into different ones.
- It is also used in classification(e.g. KNN), where the data objects are labeled based on the features' similarity.
- Another example is when we talk about dissimilar outliers compared to other data samples(e.g., anomaly detection).

Similarity and Dissimilarity Measures

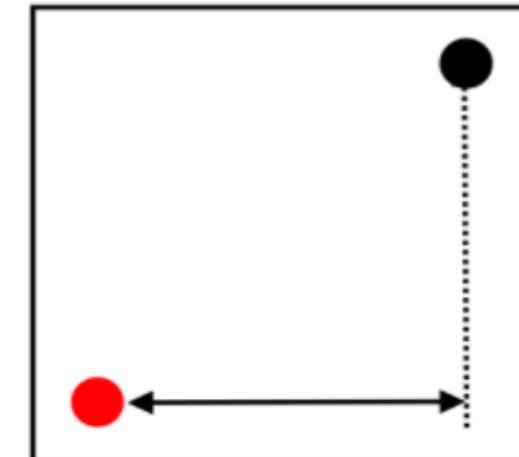
Euclidean



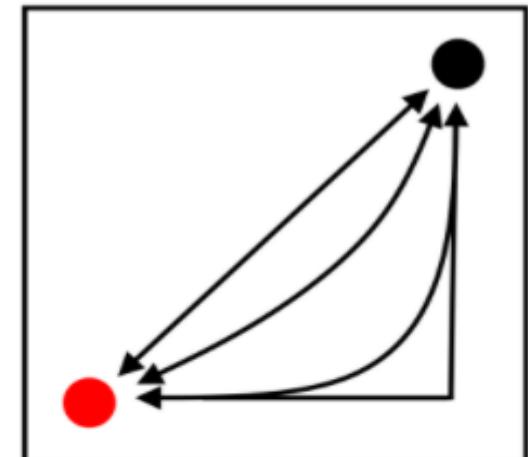
Manhattan



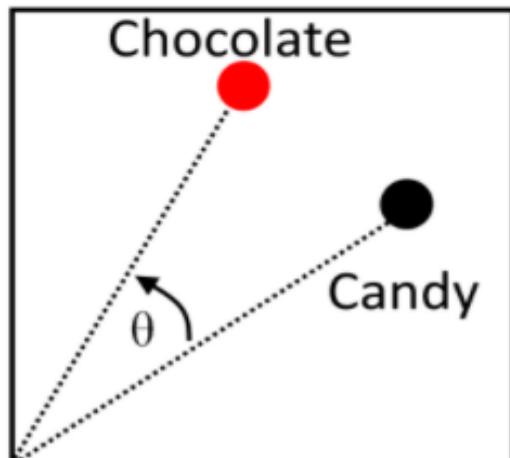
Chebychev



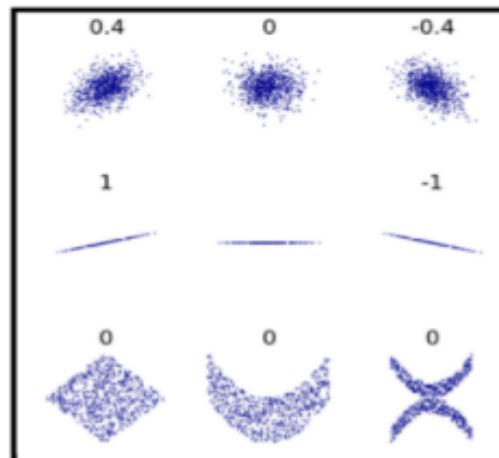
Minkowski



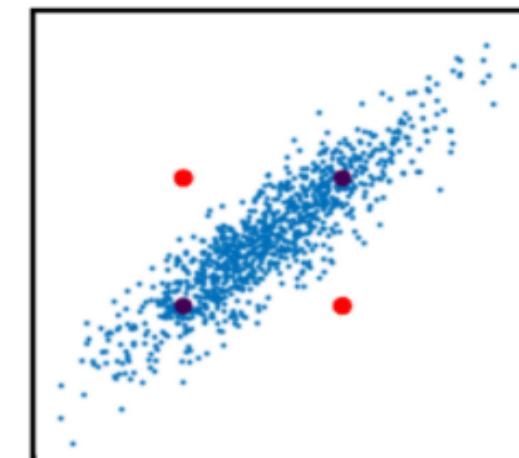
Cosine



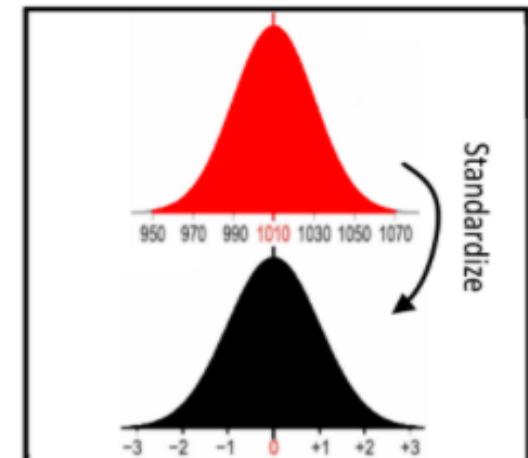
Pearson



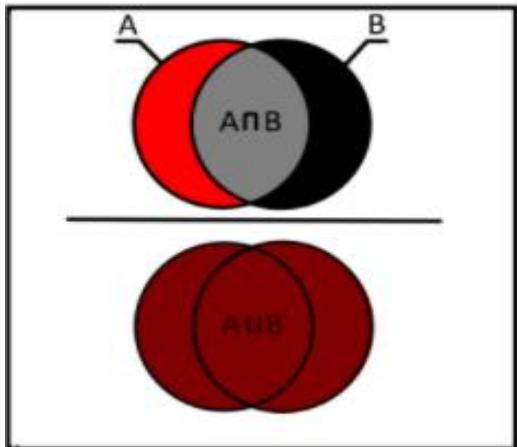
Mahalanobis



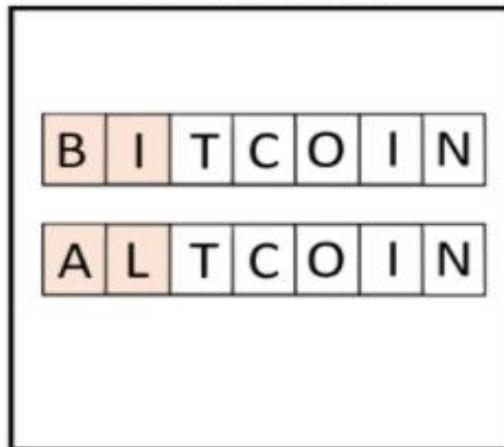
SED



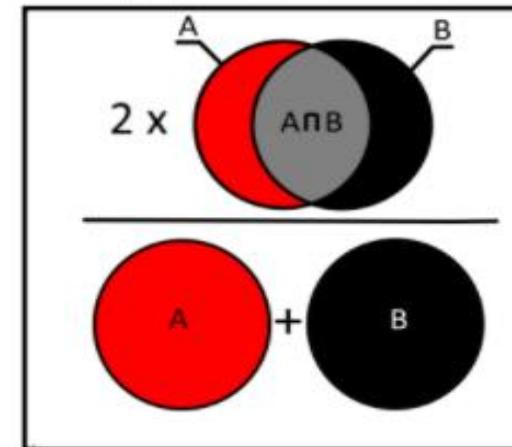
Jaccard



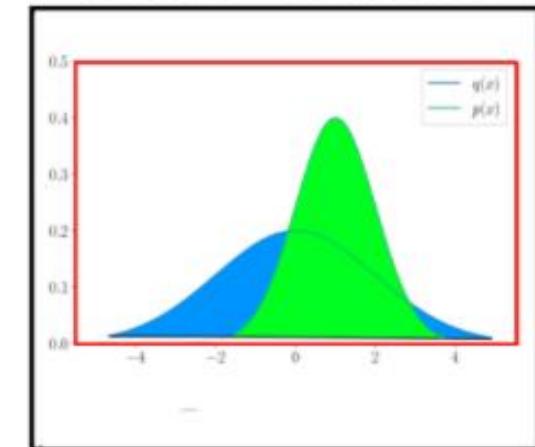
Levenshtein



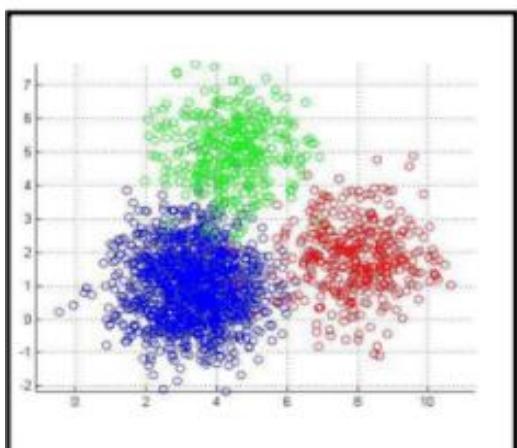
Sørensen–Dice



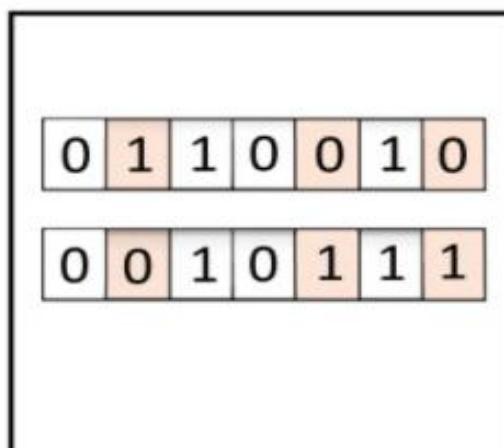
Jensen-Shannon



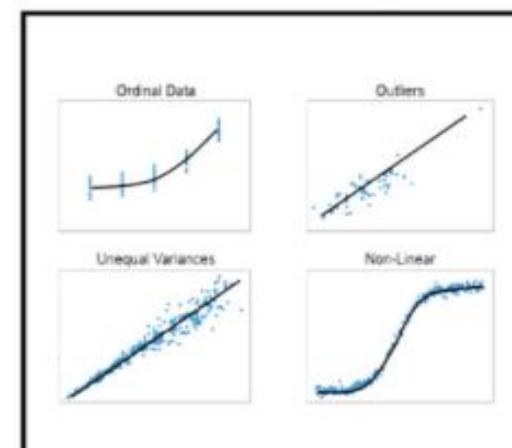
Canberra



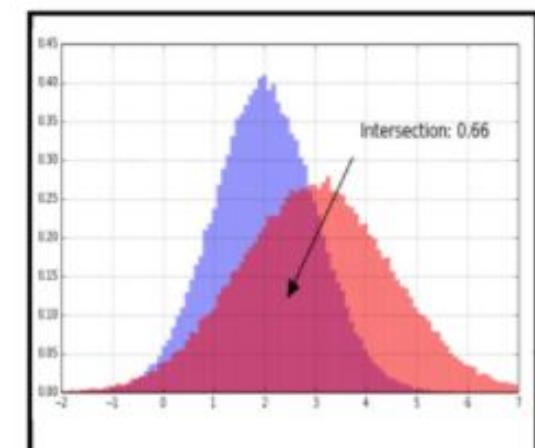
Hamming



Spearman



Chi-Square



More on Similarity Measure

- The similarity measure is usually expressed as a numerical value: It gets higher when the data samples are more alike.
- It is often expressed as a number between zero and one by conversion: zero means low similarity(the data objects are dissimilar). One means high similarity(the data objects are very similar).

Example

- Let's take an example where each data point contains only one input feature.
- This can be considered the simplest example to show the dissimilarity between three data points A, B, and C.
- Each data sample can have a single value on one axis(because we only have one input feature); let's denote that as the x-axis.
- Let's take two points, A(0.5), B(1), and C(30).
- As you can tell, A and B are close enough to each other in contrast to C.
- Thus, the similarity between A and B is higher than A and C or B and C.
- In other terms, A and B have a strong correlation. Therefore, the smaller the distance is, the larger the similarity will get

Metric:

- A given distance(e.g. dissimilarity) is meant to be a metric if and only if it satisfies the following four conditions:
 - **1- Non-negativity:** $d(p, q) \geq 0$, for any two distinct observations p and q.
 - **2- Symmetry:** $d(p, q) = d(q, p)$ for all p and q.
 - **3- Triangle Inequality:** $d(p, q) \leq d(p, r) + d(r, q)$ for all p, q, r.
 - **4-** $d(p, q) = 0$ only if $p = q$.
- Distance measures are the fundamental principle for classification, like the k-nearest neighbor's classifier algorithm, which measures the dissimilarity between given data samples.
- Additionally, choosing a distance metric would have a strong influence on the performance of the classifier.
- Therefore, the way you compute distances between the objects will play a crucial role in the classifier algorithm's performance.

Distance Functions:

- The technique used to measure distances depends on a particular situation you are working on.
- For instance, in some areas, the euclidean distance can be optimal and useful for computing distances.
- Other applications require a more sophisticated approach for calculating distances between points or observations like the cosine distance.
- The following enumerated list represents various methods of computing distances between each pair of data points.

Distance metrics

- Distance metrics are a key part of several machine learning algorithms. These distance metrics are used in both supervised and unsupervised learning, generally to calculate the similarity between data points.
- An effective distance metric improves the performance of our machine learning model, whether that's for classification tasks or clustering.

Why to learn Distance Measure?

- A distance measure is an objective score that summarizes the relative difference between two objects in a problem domain.
- Most commonly, the two objects are rows of data that describe a subject (such as a person, car, or house), or an event (such as a purchase, a claim, or a diagnosis).
- Perhaps the most likely way you will encounter distance measures is when you are using a specific machine learning algorithm that uses distance measures at its core. The most famous algorithm of this type is the k-nearest neighbors algorithm, or KNN for short.

Distance Metric Properties

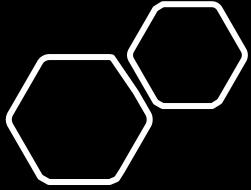
- A **metric** is a **distance function f** defined that have the following properties.
 1. **nonnegativity:** $f(x, y) \geq 0$; Distance is non-negative
 2. **reflexivity:** $f(x, y) = 0 \Leftrightarrow x = y$; Distance an object to itself is 0.
 3. **commutativity/ symmetry:** $f(x, y) = f(y, x)$; Distance is a symmetric function.
 4. **triangle inequality:** $f(x, y) \leq f(x, z) + f(y, z)$, where x , y , and z are arbitrary data points. Going directly fro object x to y in space is no more than taking detour over any object z .

A distance that satisfies these properties is a **metric**

Similarity measures for numeric data

- **Similarity measure** is the numerical measure of the degree to which two data objects are alike.
- A similarity coefficient indicates the strength of the relationship between two data points. The more the two data points resemble one another, the larger the similarity coefficient will be.
- It often fall between **0 (no similarity)** and **1 (complete similarity)**
- Similarity might be used to identify
 - duplicate data that may have differences due to typos.
 - equivalent instances from different data sets. E.g. names and/or addresses that are the same but have misspellings.
 - groups of data that are very close (**clusters**)





Similarity Measures Applications

- Clustering
- Outlier Analysis
- Nearest Neighbor Classification
- Recommendation engines
- Text related preprocessing techniques
- Different classification problems
- Email spam or ham classification problems

Dissimilarity Measure

- **Dissimilarity measure** is the numerical measure of how different two data objects are. It range from 0 (objects are alike) to ∞ (objects are different). Distance is used as a synonym for dissimilarity
- Dissimilarity might be used to identify
 - outliers
 - interesting exceptions, e.g. credit card fraud
 - boundaries to clusters
 - **Proximity Measures**
- It can be either similarity or dissimilarity.

Data matrix & Dissimilarity Matrix

Data matrix: A data matrix of n data points with l dimensions. This structure stores the n data objects in the form of a relational table, or n -by- l matrix (n objects $\times l$ attributes). Each row corresponds to an object.

$$D = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1l} \\ x_{21} & x_{22} & \dots & x_{2l} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nl} \end{pmatrix}$$

A data matrix is made up of two entities or “things”, namely rows (for objects) and columns (for attributes). Therefore, the data matrix is often called a two-mode matrix. The dissimilarity matrix contains one kind of entity (dissimilarities) and so is called a one-mode matrix.

- **Dissimilarity matrix** is a triangular matrix of n data points that registers only the distance of dissimilarity. It stores a collection of proximities for a pair of n objects. $d(i, j)$ is the measured dissimilarity or “difference” between objects i and j .

$$\begin{bmatrix} d(2,1) & 0 \\ d(3,1) & d(3,2) & 0 \\ \vdots & \vdots & \vdots \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$



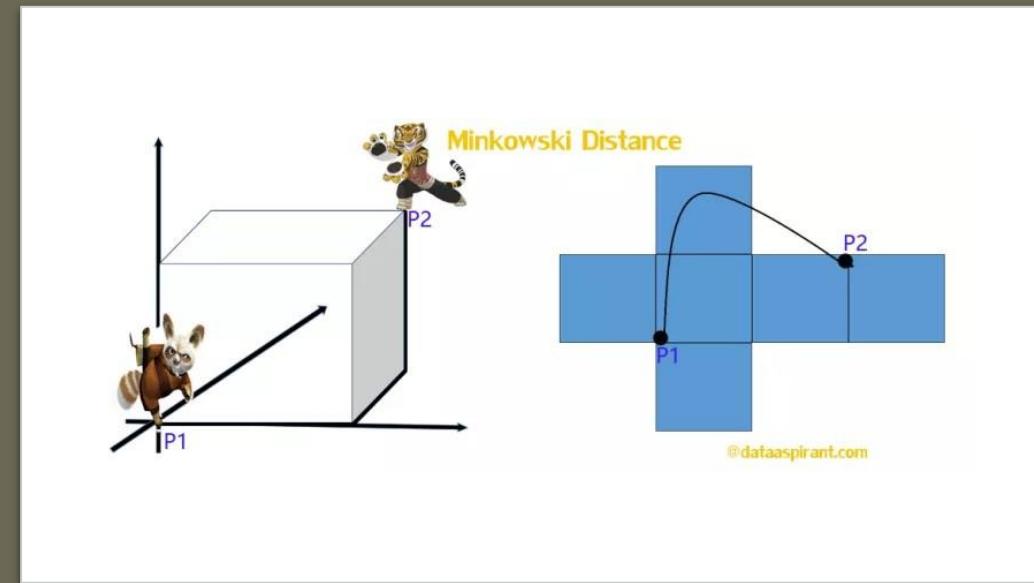
Dissimilarity of Numeric Data

1. Minkowski distance
2. Euclidean distance
3. Manhattan distance
4. Supremum distance
5. Mahalanobis distance
6. Bhattacharyya distance

Minkowski Distance

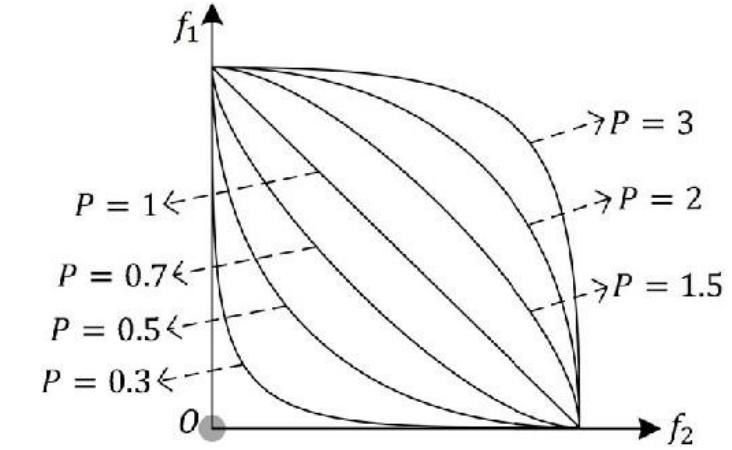
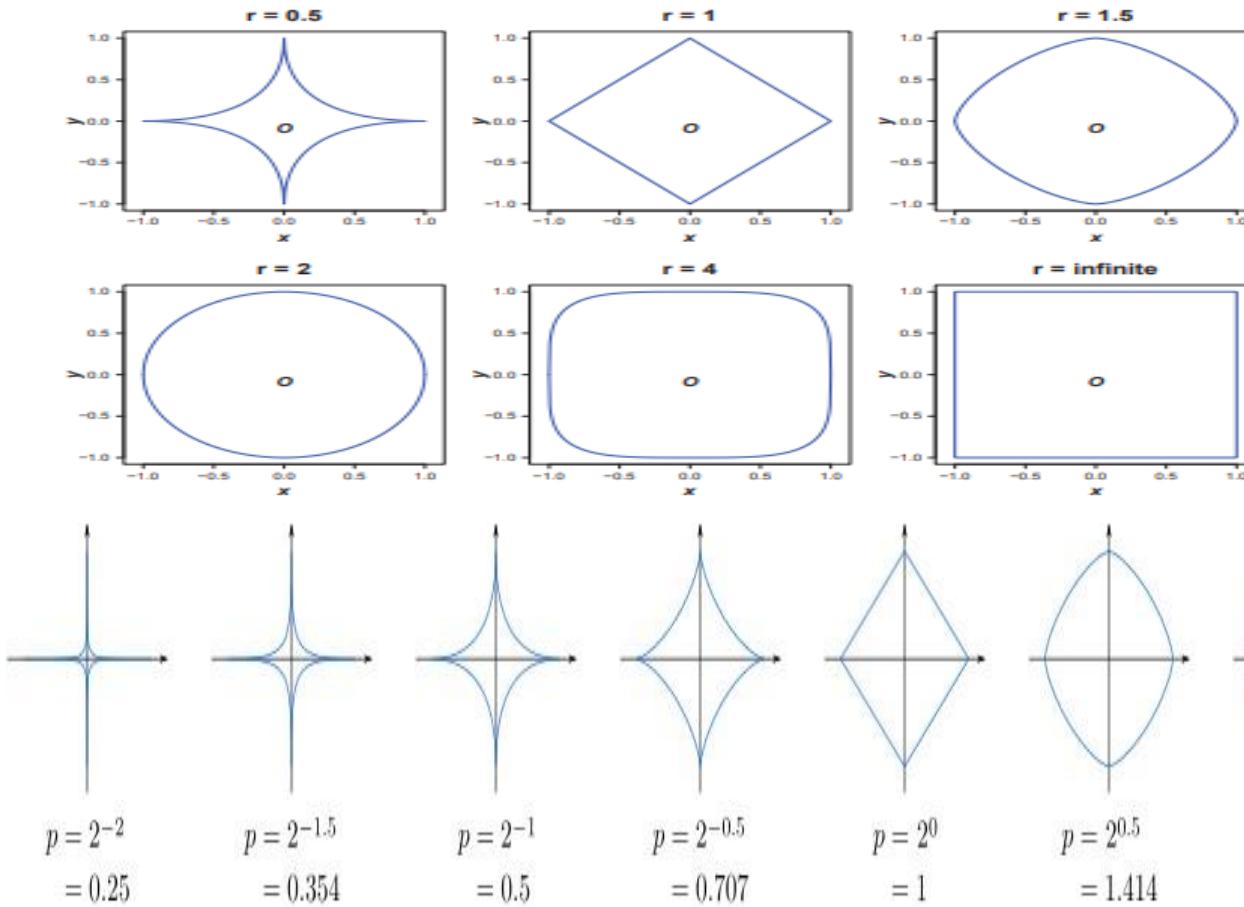
1. Minkowski Distance

- Minkowski Distance is a generalization of Euclidean Distance and Manhattan distance in a normed vector space.
- Normed vector space. What is Normed vector space? A Normed vector space is a vector space on which a norm is defined. Suppose X is a vector space then a norm on X is a real valued function $\|x\|$ which satisfies below conditions -
 - **Zero Vector-** Zero vector will have zero length.
 - **Scalar Factor-** The direction of vector doesn't change when you multiply it with a positive number though its length will be changed.
 - **Triangle Inequality-** If distance is a norm then the calculated distance between two points will always be a straight line.
- Although p can be any real value, it is typically set to a value between 1 and 2. For values of p less than 1, the formula above does not define a valid distance metric since the triangle inequality is not satisfied.



$$D(X, Y) = \left(\sum_{i=1}^n |x_i - y_i|^p \right)^{\frac{1}{p}}$$

Minkowski distance



Example of different Minkowski distance contour curves when the value of P alters

- The following figure shows unit circles (the set of all points that are at the unit distance from the centre) with various values of p . All the points that are at a distance of 1 from the center, which is the definition of a circle in Euclidean distance ($r = 2$). Notice how that circle grows progressively until reaching the square form in the infinity ($r \rightarrow +\infty$). This is because when r increases, the influence of the highest component $|x_i - y_i|$ r in equation increases notably compared to the other components in the distance computation

Synonyms of Minkowski

- Different names for the Minkowski distance or Minkowski metric arise from the order:
- $P=1$ is the **Manhattan distance**. Synonyms are **L1-Norm**, **Taxicab**, or **City-Block distance**. For two vectors of ranked ordinal variables, the Manhattan distance is sometimes called **Foot-ruler distance**.
- $P=2$ is the **Euclidean distance**. Synonyms are **L2-Norm** or **Ruler distance**. For two vectors of ranked ordinal variables, the Euclidean distance is sometimes called **Spear-man distance**.
- $P=\infty$ is the **Chebyshev distance**. Synonyms are **Lmax-Norm** or **Chessboard distance**.

Example Minkowski Distance

Features	Coord1	Coord2	Coord3	Coord4	Coord5	Coord6
Object A	0	3	4	5		
Object B	7	6	3	-1		

Minkowski distance for order 3 is

$$\begin{aligned}d_{BA} &= \left(|0 - 7|^3 + |3 - 6|^3 + |4 - 3|^3 + |5 + 1|^3 \right)^{\frac{1}{3}} \\&= \sqrt[3]{343 + 27 + 1 + 216} = \sqrt[3]{587} = 8.373\end{aligned}$$

Example Minkowski Distance

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

The Manhattan distance is obtained setting p=1 in the Minkowski distance

L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

The Euclidean distance is obtained setting p=2 in the Minkowski distance

L2	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

The Chebyshev distance is obtained setting p=inf in the Minkowski distance

L ∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

Distance Matrix

Disadvantages Minkowski

- Minkowski has the same disadvantages as the distance measures they represent, so a good understanding of metrics like Manhattan, Euclidean, and Chebyshev distance is extremely important.
- Moreover, the parameter p can actually be troublesome to work with as finding the right value can be quite computationally inefficient depending on your use-case.
- Use Cases
- The upside to p is the possibility to iterate over it and find the distance measure that works best for your use case. It allows you a huge amount of flexibility over your distance metric, which can be a huge benefit if you are closely familiar with p and many distance measures.

Euclidean distance

Euclidian Distance

- In mathematics, the **Euclidean distance** between two points in Euclidean space is the length of a line segment between the two points.
- It can be calculated from the Cartesian coordinates of the points using the Pythagorean theorem, therefore occasionally being called the **Pythagorean distance**.
- These names come from the ancient Greek mathematicians Euclid and Pythagoras, although Euclid did not represent distances as numbers, and the connection from the Pythagorean theorem to distance calculation was not made until the 18th century.
- The distance between two objects that are not points is usually defined to be the smallest distance among pairs of points from the two objects.
- Formulas are known for computing distances between different types of objects, such as the distance from a point to a line.
- In advanced mathematics, the concept of distance has been generalized to abstract metric spaces, and other distances than Euclidean have been studied.
- In some applications in statistics and optimization, the square of the Euclidean distance is used instead of the distance itself.

History of Euclidian Distance

- Euclidean distance is the distance in Euclidean space; both concepts are named after ancient Greek mathematician Euclid, whose *Elements* became a standard textbook in geometry for many centuries.
- Concepts of length and distance are widespread across cultures, can be dated to the earliest surviving "protoliterate" bureaucratic documents from Sumer in the fourth millennium BC (far before Euclid), and have been hypothesized to develop in children earlier than the related concepts of speed and time.
- But the notion of a distance, as a number defined from two points, does not actually appear in Euclid's *Elements*. Instead, Euclid approaches this concept implicitly, through the congruence of line segments, through the comparison of lengths of line segments, and through the concept of proportionality.

History of Euclidian Distance

- The Pythagorean theorem is also ancient, but it could only take its central role in the measurement of distances after the invention of Cartesian coordinates by René Descartes in 1637.
- The distance formula itself was first published in 1731 by Alexis Clairaut because of this formula, Euclidean distance is also sometimes called Pythagorean distance.
- Although accurate measurements of long distances on the earth's surface, which are not Euclidean, had again been studied in many cultures since ancient times (see history of geodesy), the idea that Euclidean distance might not be the only way of measuring distances between points in mathematical spaces came even later, with the 19th-century formulation of non-Euclidean geometry.
- The definition of the Euclidean norm and Euclidean distance for geometries of more than three dimensions also first appeared in the 19th century, in the work of Augustin-Louis Cauchy.

Euclidian Distance Formula

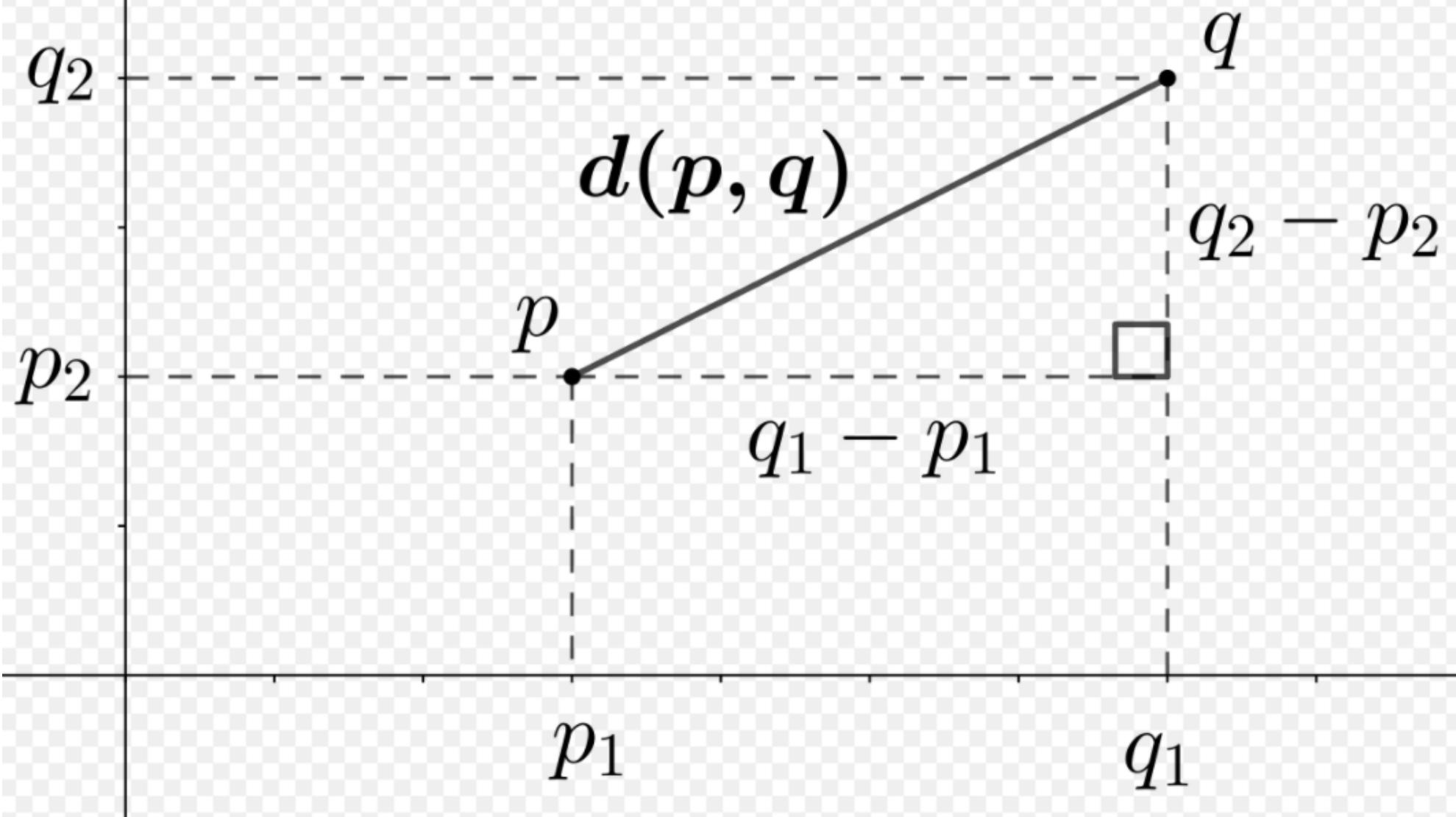
It is a distance measure that best can be explained as the length of a segment connecting two points.

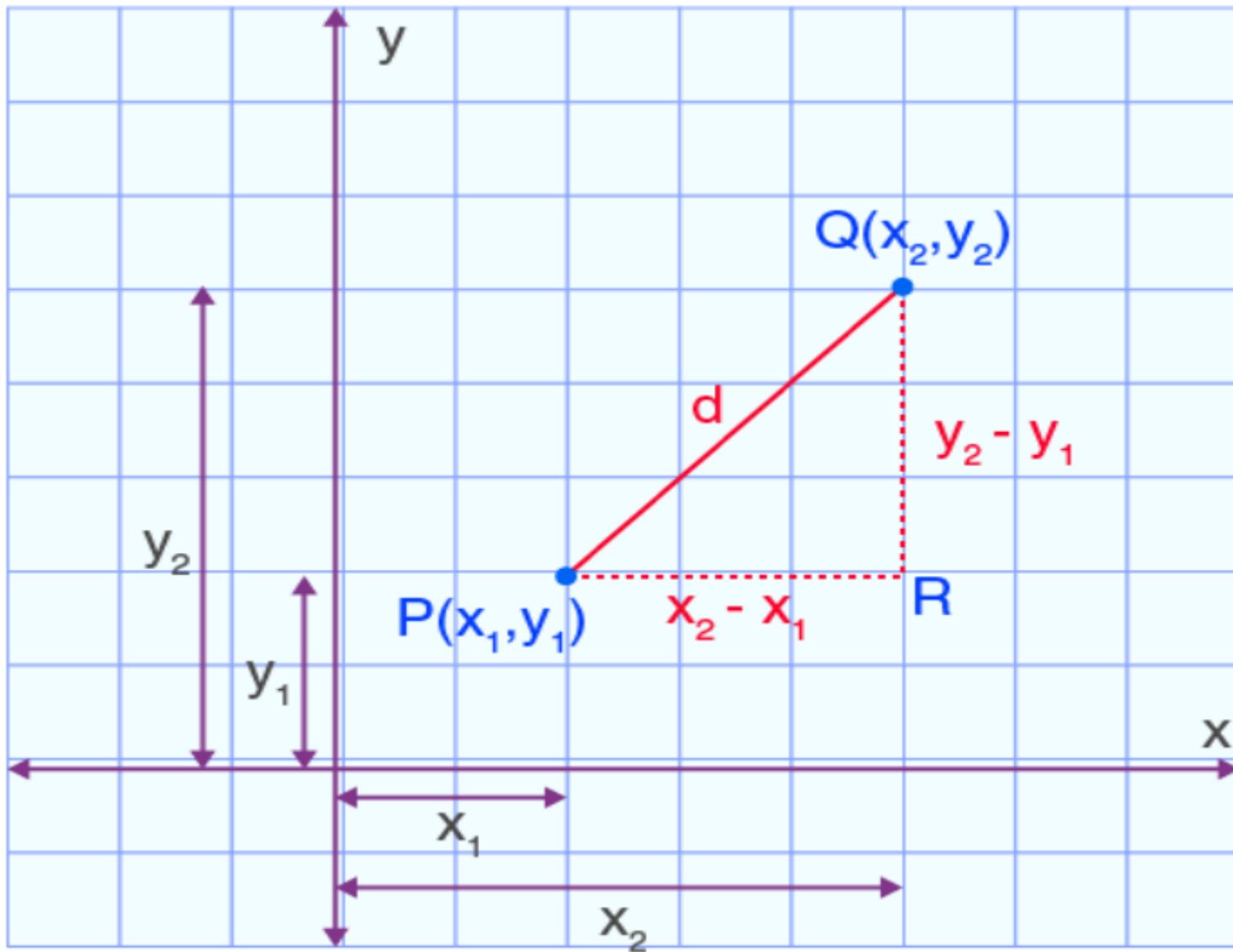
The formula is rather straightforward as the distance is calculated from the cartesian coordinates of the points using the Pythagorean theorem

$$D(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

Euclidean distance

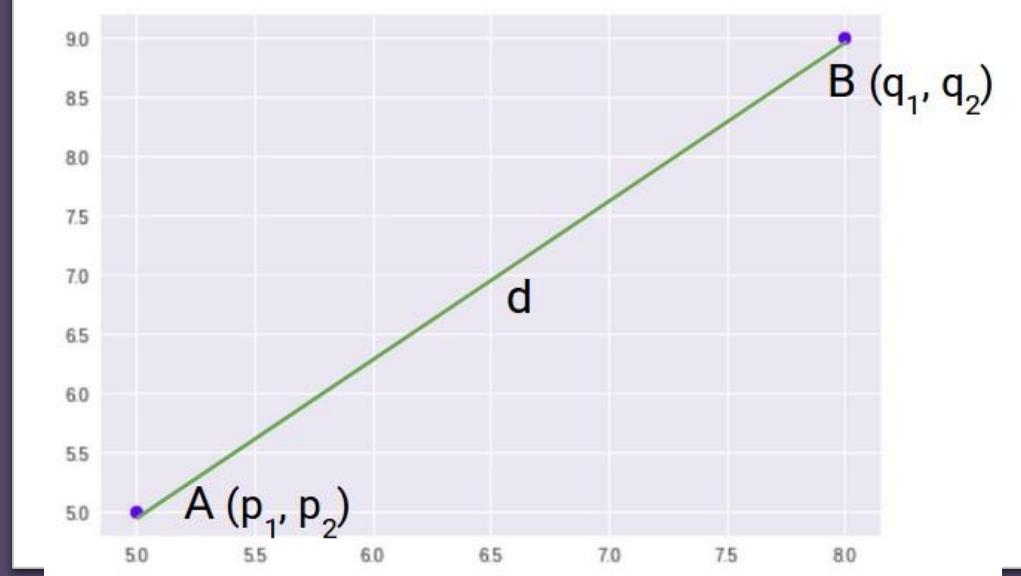
$$d(p, q)^2 = (q_1 - p_1)^2 + (q_2 - p_2)^2$$





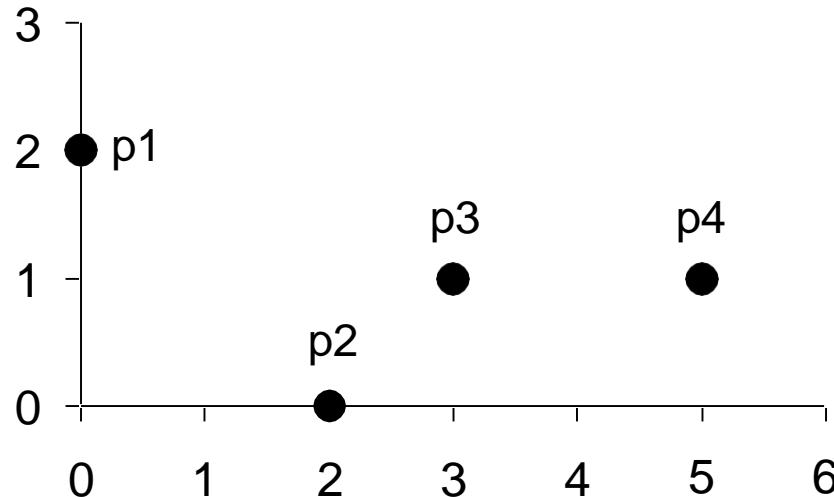
Summary: Euclidean distance

- The Euclidean distance between two points (x, y) in any dimension of space is the length of the path connecting them. The **Pythagorean theorem** gives this distance between two points.
- **Euclidean Distance** represents the shortest distance between two points
- When data is **dense or continuous**, Euclidean distance is the best proximity measure.
- where n is the number of dimensions (attributes) and x_k and y_k are, respectively, the k^{th} attributes (components) or data objects \mathbf{x} and \mathbf{y} .
- Eg: Calculate Euclidean distance between points $[0,3,4,5], [7,6,3,-1]$



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

Euclidean Distance Examples



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

$$\sqrt{(2-0)^2 + (0-2)^2} = \sqrt{8}$$

$$\sqrt{(3-0)^2 + (1-2)^2} = \sqrt{10}$$

$$\sqrt{(5-0)^2 + (1-2)^2} = \sqrt{26}$$

	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

How to derive Euclidian Formula

- Now, we have to apply Pythagoras theorem to the triangle PQR to find the distance between two points.
- Using Pythagoras theorem,
- Hypotenuse² = Base² + Perpendicular²
- $PQ^2 = PR^2 + QR^2$
- Therefore, $d^2 = (x_2 - x_1)^2 + (y_2 - y_1)^2$
- Now, take square root on both sides of equation, we get
- $d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$
- Hence, the formula for Euclidean distance is derived.

Example

- Determine the Euclidean distance between two points (a, b) and $(-a, -b)$.
- Solution:
 - Let the point P be (a, b) and Q be $(-a, -b)$
 - i.e. $P(a, b) = (x_1, y_1)$ and $Q(-a, -b) = (x_2, y_2)$
 - We know that the Euclidean distance formula is,
 - Euclidean distance, $d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

Solution- Euclidian Distance Computation

- Now, substitute the values in the formula, we get
- $d = \sqrt{(-a - a)^2 + (-b - b)^2}$
- $d = \sqrt{(-2a)^2 + (-2b)^2}$
- $d = \sqrt{4a^2 + 4b^2}$
- $d = \sqrt{4(a^2+b^2)}$
- $d = 2\sqrt{(a^2+b^2)}$.
- Hence, the distance between two points (a, b) and $(-a, -b)$ is $2\sqrt{(a^2+b^2)}$.

Another example

- Find the distance between two points $P(0, 4)$ and $Q(6, 2)$.

Solution

- Given:

- $P(0, 4) = (x_1, y_1)$
- $Q(6, 2) = (x_2, y_2)$

The distance between the point PQ is

- $PQ = \sqrt{[(x_2 - x_1)^2 + (y_2 - y_1)^2]}$

Computation Result:

- $PQ = \sqrt{[(6 - 0)^2 + (2 - 4)^2]}$
- $PQ = \sqrt{[6^2 + (-2)^2]}$
- $PQ = \sqrt{36+4}$
- $PQ = \sqrt{40}$
- $PQ = 2\sqrt{10}.$
- Therefore, the distance between two points P(0,4) and Q(6, 2) is $2\sqrt{10}.$

Lecture 16: Week 6_

17/2/2021

Agenda

1. Manhattan distance
2. Supremum distance
3. Mahalanobis distance
4. Bhattacharyya distance

Manhattan distance

Manhattan Distance

The *Manhattan distance* between two vectors (city blocks) is equal to the one-norm of the distance between the vectors.

The distance function (also called a “metric”) involved is also called the “taxi cab” metric.

.

3. Manhattan distance

- Manhattan distance is a metric in which the distance between two points is calculated as the **sum of the absolute differences of their Cartesian coordinates**.
- Manhattan distance also known as **city block distance**, is **the distance in blocks between any 2 points**.
- Manhattan distance is usually preferred over the more common Euclidean distance when there is **high dimensionality in the data**.
- In a simple way of saying it is the total sum of the difference between the x-coordinates and y-coordinates.
- Manhattan distance = $|x_1 - x_2| + |y_1 - y_2|$

$$D_m = \sum_{i=1}^n |p_i - q_i|$$

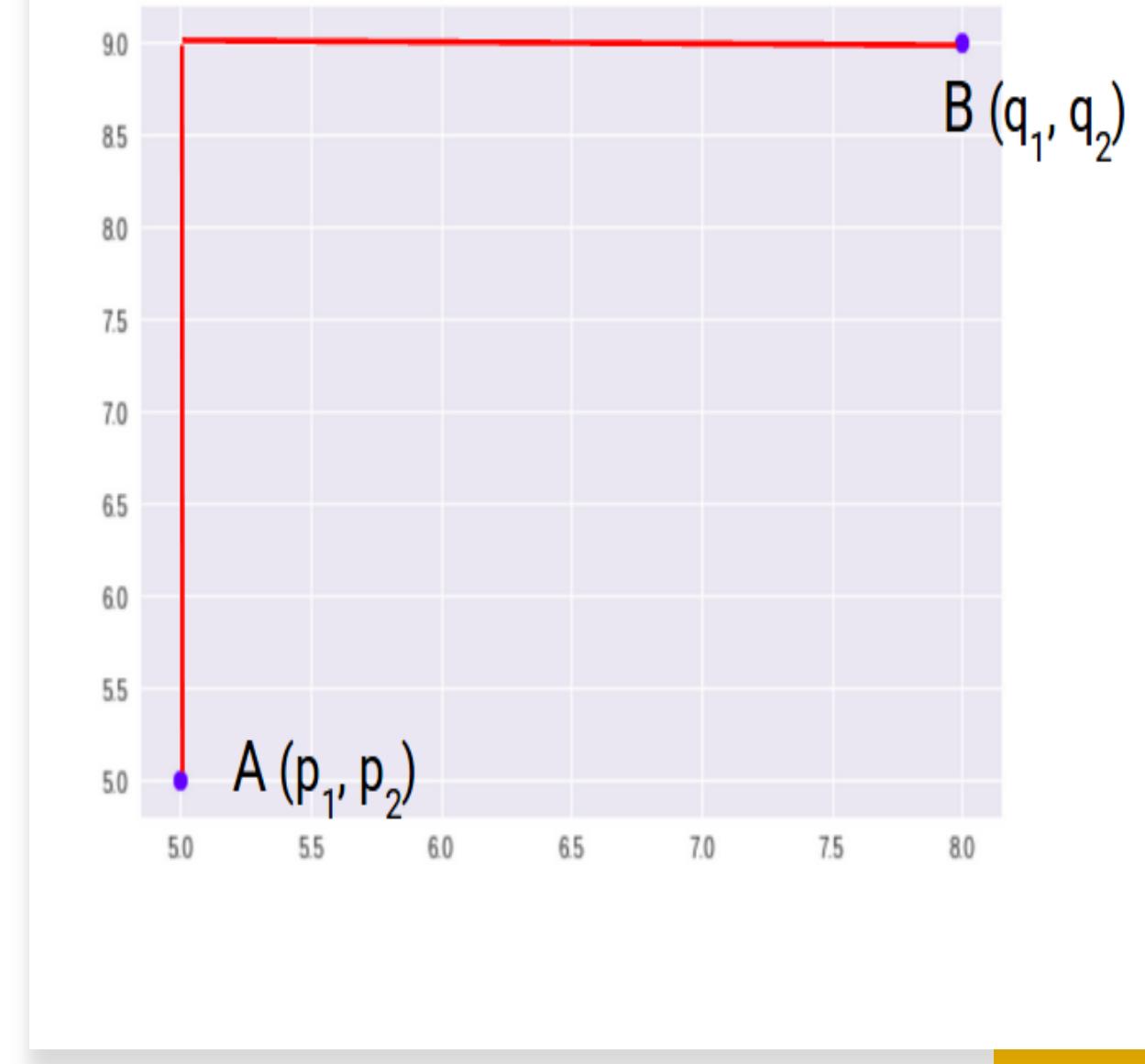


Illustration of Manhattan Distance

The Manhattan distance as the sum of absolute differences

Manhattan Distance

$$[\{a, b, c\}, \{x, y, z\}]$$

$$\rightarrow \text{Abs } [a - x] + \text{Abs } [b - y] + \text{Abs } [c - z]$$

EXAMPLE OF MANHATTAN DISTANCE

The one-norm as Manhattan distance between two city blocks

$$\text{block1} = \{1, 2, 3, 4\};$$

$$\text{block2} = \{5, 6, 7, 8\};$$

$$\text{Norm} [\text{block1} - \text{block2}, 1]$$

$$= |1 - 5| + |2 - 6| + |3 - 7| + |8 - 4|$$

$$= 16$$



The Manhattan length of Two Blocks

The Manhattan length of two blocks

block1 = {5, 2, -3, 4};
block2 = {1, 6, -7, 8};

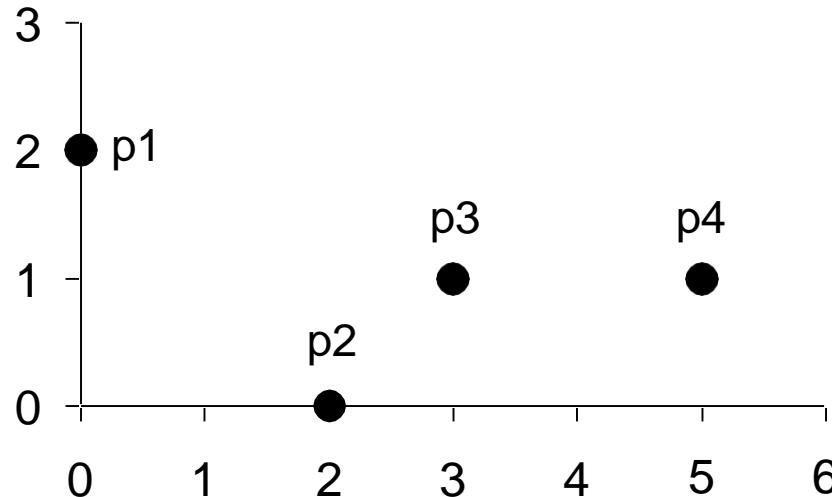
→ { Norm [block1, 2], Norm [block2, 1] }

→ $|5| + |2| + |-3| + |4| = 14$

→ $|1| + |6| + |-7| + |8| = 22$

{14, 22}

Additional Manhattan Distance Examples



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

$$|2-0| + |0-2| = 4$$

$$|3-0| + |1-0| = 4$$

$$|5-0| + |1-2| = 6$$

	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

Distance Matrix

Supremum Distance

Supremum Distance

- The supremum distance (also referred to as L_{\max} , L_∞ norm and as the Chebyshev distance) is a **generalization of the Minkowski distance for $h \rightarrow \infty$** .
- To compute it, we find the attribute f that gives the maximum difference in values between the two objects.

Definition of Supremum Distance

This difference is the supremum distance, defined more formally as:

$$d(i, j) = \lim_{h \rightarrow \infty} \left(\sum_{f=1}^p |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_f |x_{if} - x_{jf}|.$$

The L^∞ norm is also known as the *uniform norm*.

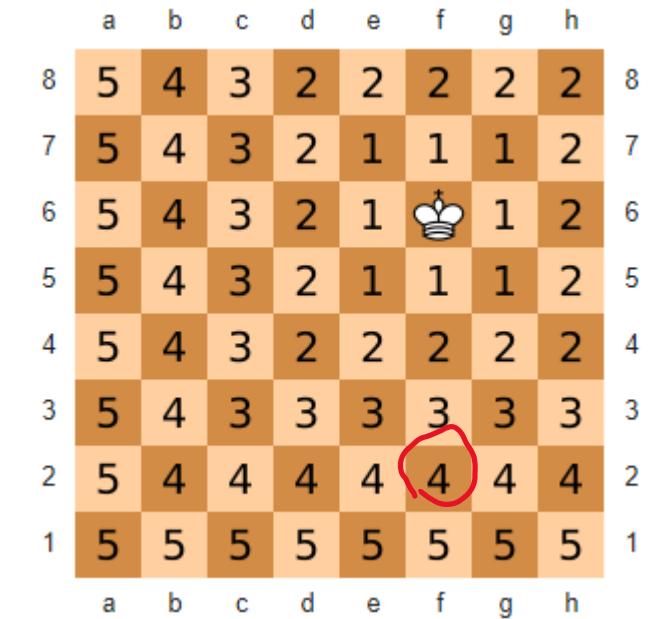
Chebyshev distance

- In mathematics, **Chebyshev distance** (or **Tchebychev distance**), **maximum metric**, or L_∞ metric is a metric defined on a vector space where the distance between two vectors is the greatest of their differences along any coordinate dimension. It is named after **Pafnuty Chebyshev**.

Chessboard distance

This Chebyshev problem is called chess board distance problem since in the game of [chess](#) the minimum number of moves needed by a [king](#) to go from one square on a [chessboard](#) to another equals the Chebyshev distance between the centers of the squares, if the squares have side length one, as represented in 2-D spatial coordinates with axes [Rajeshgned](#) to the edges of the board.

For example, the Chebyshev distance between f6 and e2 equals 4.



Supremum Distance

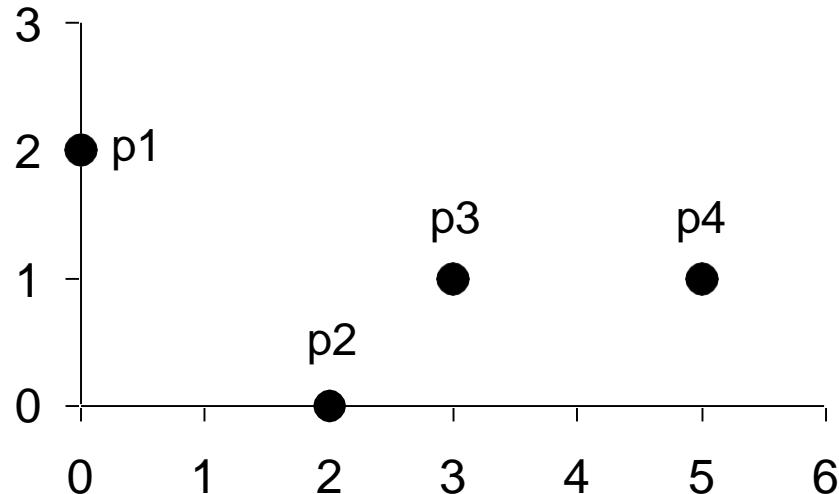
- Supremum Distance is **the maximum difference between any component** (attribute) of the vectors. It is a metric defined on a vector space **where distance between two vectors is the greatest of their difference along any coordinate dimension.** $CD(x, y) = \max_i |x_i - y_i|$
- Synonyms are **L_{max}-Norm** or **Chessboard distance.** P= ∞ on Minkowski distance is the Chebyshev distance also known as supremum distance.
- Chebyshev distance is appropriate in cases when two objects are to be defined as different **if they are different in any one dimension.**

$$d(i, j) = \lim_{h \rightarrow \infty} \left(\sum_{f=1}^p |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_f |x_{if} - x_{jf}|$$

$$\begin{aligned} d_{BA} &= \max(|0-7|, |3-6|, |4-3|, |5+1|) \\ &= \max\{7, 3, 1, 6\} = 7 \end{aligned}$$

Features	Coord1	Coord2	Coord3	Coord4	Coord5	Coord6
Object A	0	3	4	5		
Object B	7	6	3	-1		

Chebyshev Distance Examples



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

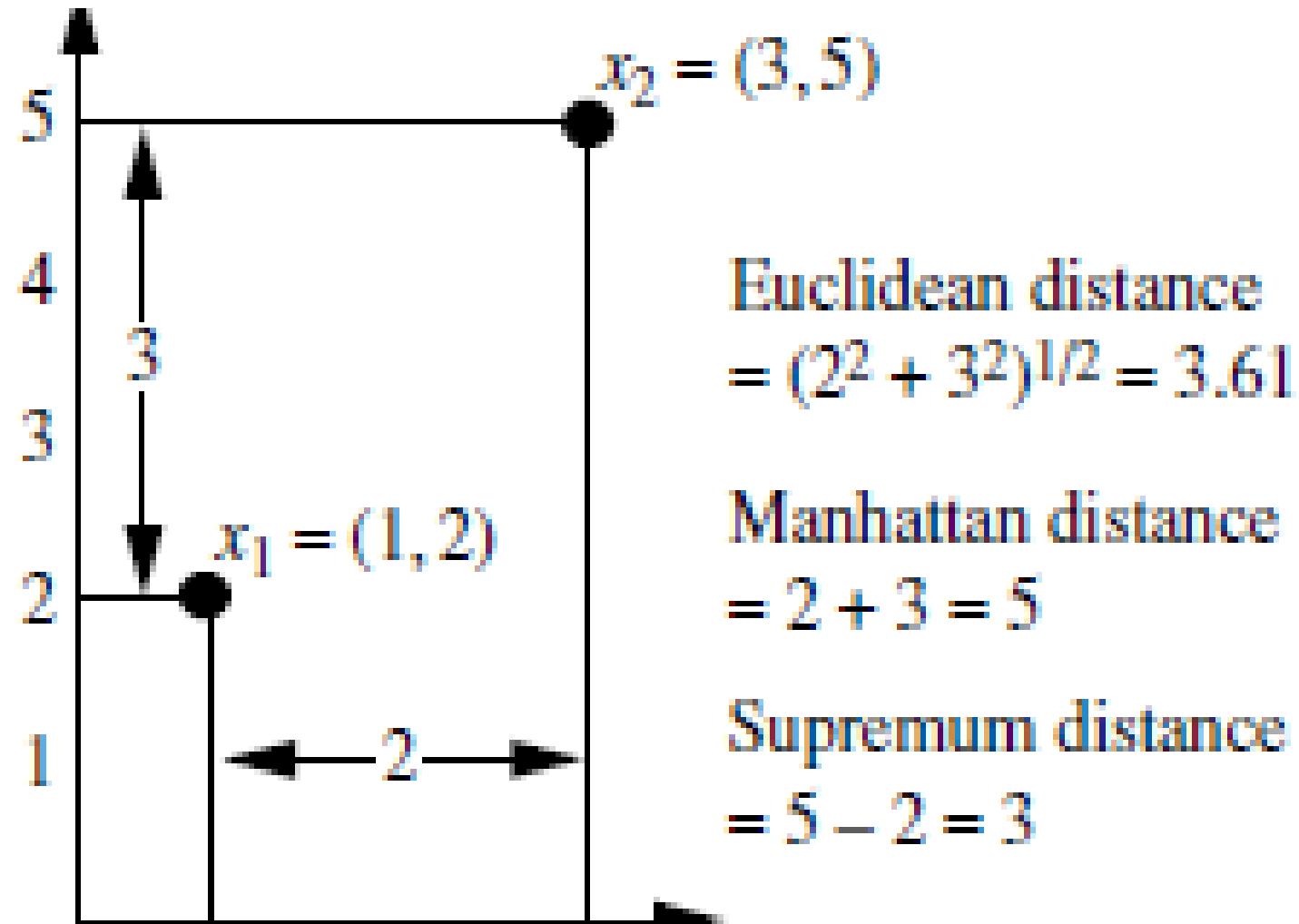
$$\text{Max}(|2-0|, |0-2|) = 2$$

$$\text{Max}(|3-0|, |1-0|) = 3$$

$$\text{Max}(|5-0|, |1-2|) = 5$$

L_∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

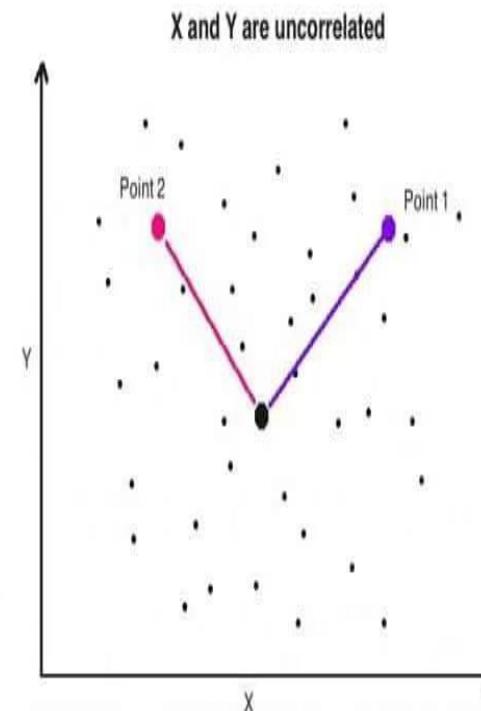
Distance Matrix



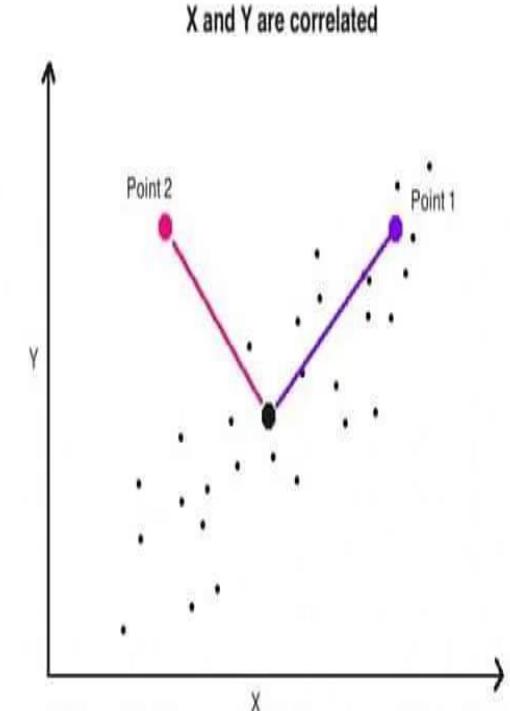
Manhattan,
Euclidean &
Supremum
Distance
between two
objects

Euclidean distance Issues

- The two points above are equally distant (Euclidean) from the center. But only one of them (blue) is actually more close to the cluster, even though, technically the Euclidean distance between the two points are equal.
- This is because, Euclidean distance is a distance between two points only. It does not consider how the rest of the points in the dataset vary. So, it cannot be used to really judge how close a point actually is to a distribution of points.



When X and Y are uncorrelated, the Euclidean distance from the Centroid can be useful to infer if a point is member of the distribution. The farther it is, the less likely it is a member.



Both Point 1 and Point 2 have the same Euclidean distance from centroid. But only Point 1 is a member of the distribution. To detect Point 2 as outlier, `dist(Point 2, centroid)` should be much higher than `dist(Point 1, centroid)`. Mahalanobis distance can be used here instead.

Euclidean distance Issues

- The two tables above show the ‘area’ and ‘price’ of the same objects. Only the units of the variables change.
- Since both tables represent the same entities, the distance between any two rows, point A and point B should be the same. But Euclidean distance gives a different value even though the distances are technically the same in physical space.

Area (sq.ft)	Price (\$ 1000's)	Area (acre)	Price (\$M)
2400	156000	0.0550944	156
1950	126750	0.0447642	126.75
2100	105000	0.0482076	105
1200	78000	0.0275472	78
2000	130000	0.045912	130
900	54000	0.0206604	54

Mahalanobis Distance

Mahalanobis Distance

- The **Mahalanobis distance** is a measure of the distance between a point P and a distribution D , introduced by P. C. Mahalanobis in 1936.
- It is a multi-dimensional generation of the idea of measuring how many standard deviations away P is from the mean of D .
- This distance is zero for P at the mean of D and grows as P moves away from the mean along each principal component axis.
- If each of these axes is re-scaled to have unit variance, then the Mahalanobis distance corresponds to standard Euclidean distance in the transformed space.
- The Mahalanobis distance is thus unitless, scale-invariant, and takes into account the correlations of the data set.

Definition

- The Mahalanobis distance of an observation $\vec{x} = (x_1, x_2, x_3, \dots, x_N)^\top$ from a set of observations with mean $\vec{\mu} = (\mu_1, \mu_2, \mu_3, \dots, \mu_N)^\top$ and (nonsingular) covariance matrix \mathbf{S} is defined as

$$D_M(\vec{x}) = \sqrt{(\vec{x} - \vec{\mu})^\top \mathbf{S}^{-1} (\vec{x} - \vec{\mu})}.$$

Applications of Mahalanobis Distance

- Problem of identifying the similarities of skulls based on measurements in 1927
- Mahalanobis distance is widely used in cluster analysis and classification techniques.
- It helps in multivariate statistical testing and Fisher's Linear Discriminant Analysis that is used for supervised classification.
- Its shows leverage to detect outliers and have greater influence on the slope or coefficients of the regression equation of linear regression models.
- Mahalanobis distance is also used to determine multivariate outliers.
- Even for normal distributions, a point can be a multivariate outlier, Mahalanobis distance a more sensitive measure than checking dimensions individually.,
- Mahalanobis distance has also been used in ecological niche modelling.

Mahalanobis distance.

- It is a metric of measure mostly used in multivariate statistical testing where the euclidean distance fails to give the real distance between observations.
- It measures how far away a data point from the distribution

Mahalanobis Distance

- Mahalanobis Distance is a measure of distance between a **data vector and a set of data**, or a variation that measures the **distance between two vectors from the same dataset**.
- Mahalanobis Distance is used for calculating the distance between two data points in a **multivariate space**
- Also known as **quadratic distance**, measures separation of 2 groups of objects.

$$\text{mahalanobis } (\mathbf{x}, \mathbf{y}) = ((\mathbf{x} - \mathbf{y})^T \Sigma^{-1}(\mathbf{x} - \mathbf{y}))^{0.5}$$

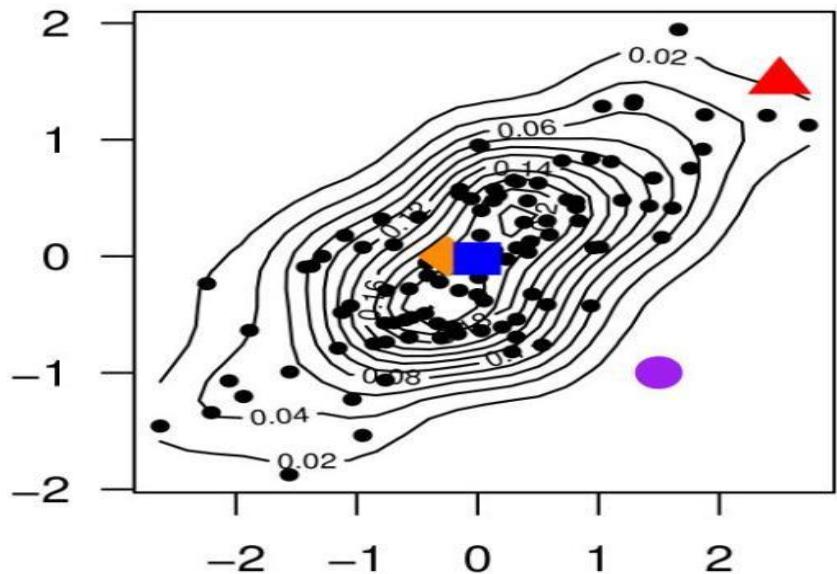
Σ is the covariance matrix

- The data of both group should have same number of variables(no. Of columns) but not same number of data(rows).
- *The Mahalanobis distance is a measure of the distance between a point P and a distribution D. The idea of measuring is, how many standard deviations away P is from the mean of D.*
- The benefit of using mahalanobis distance is, it takes **covariance** in account which helps in measuring the strength/similarity between two different data objects.
- **When the covariance matrix is identity Matrix, the mahalanobis distance is the same as the Euclidean distance.**
- **Useful for detecting outliers(multivariate), multivariate anomaly detection, classification** on highly imbalanced datasets and one-class classification

Formulas

$$D_M(\vec{x}) = \sqrt{(\vec{x} - \vec{\mu})^T S^{-1} (\vec{x} - \vec{\mu})}.$$

$$d(\vec{x}, \vec{y}) = \sqrt{(\vec{x} - \vec{y})^T S^{-1} (\vec{x} - \vec{y})}.$$



$$d_M(x, y) = \sqrt{(x - y)^T S^{-1} (x - y)}$$

$$= \sqrt{[x_1 - y_1 \quad x_2 - y_2] \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{bmatrix} [x_1 - y_1 \quad x_2 - y_2]}$$

$$= \sqrt{\left[\frac{x_1 - y_1}{\sigma_1^2} \quad \frac{x_2 - y_2}{\sigma_2^2} \right] [x_1 - y_1 \quad x_2 - y_2]}$$

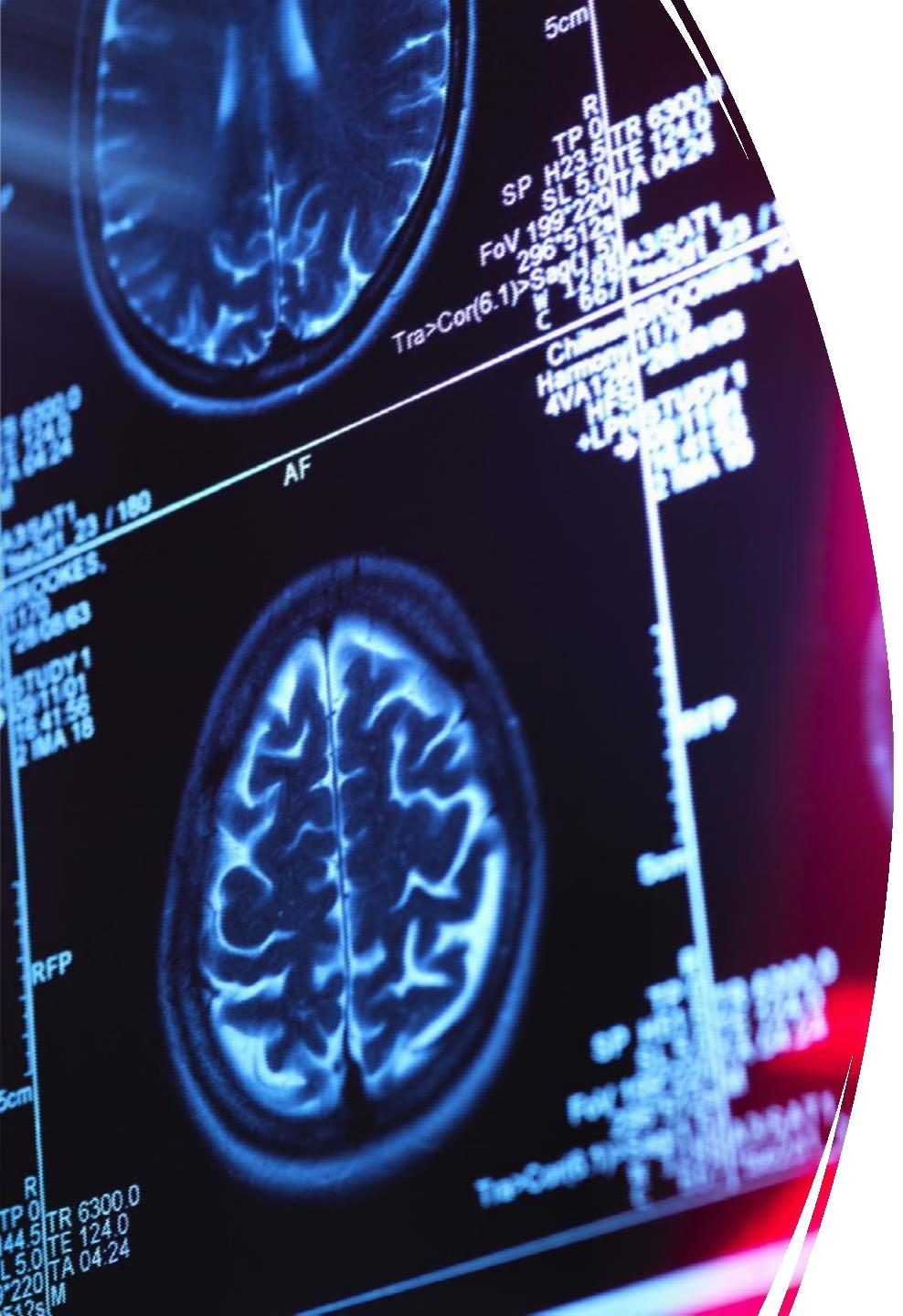
$$= \sqrt{\frac{(x_1 - y_1)^2}{\sigma_1^2} + \frac{(x_2 - y_2)^2}{\sigma_2^2}}$$

Example

X	Y	Z
Height	Score	Age
64.0	580.0	29.0
66.0	570.0	33.0
68.0	590.0	37.0
69.0	660.0	46.0
73.0	600.0	55.0

$$m = 68.0 \quad 600.0 \quad 40.0$$

$n=5$. you want to know how far another person, $v = (66, 640, 44)$, is from this data



COVARIANCE MATRIX

$$\text{COV}(X,Y,Z) = \begin{vmatrix} \text{cov}(X,X) & \text{cov}(X,Y) & \text{cov}(X,Z) \\ \text{cov}(Y,X) & \text{cov}(Y,Y) & \text{cov}(Y,Z) \\ \text{cov}(Z,X) & \text{cov}(Z,Y) & \text{cov}(Z,Z) \end{vmatrix}$$

SOLUTION

X	Y	Z		X-Xbar	Y-Ybar	Z-Zbar		(X-Xbar)^2	(Y-Ybar)^2	(Z-Zbar)^2	
64	580	29		-4	-20	-11		16	400	121	
66	570	33		-2	-30	-7		4	900	49	
68	590	37		0	-10	-3		0	100	9	
69	660	46		1	60	6		1	3600	36	
73	600	55		5	0	15		25	0	225	
Sum	340	3000	200	n=5				SUM	46	5000	440
Mean	68	600	40	n-1=4				VARIANCE	11.5	1250	110

Bhattacharyya Distance

Bhattacharyya Distance

- In [statistics](#), the **Bhattacharyya distance** [measures the similarity](#) of two [probability distributions](#).
- It is closely related to the **Bhattacharyya coefficient** which is a measure of the amount of overlap between two [statistical](#) samples or populations.
- Both measures are named after [Anil Kumar Bhattacharyya](#), a [statistician](#) who worked in the 1930s at the [Indian Statistical Institute](#).
- He has developed the method to measure the distance between two non-normal distributions and illustrated this with the classical multinomial populations^[1] as well as probability distributions that are absolutely continuous with respect to the Lebesgue measure.
- The latter work appeared partly in 1943 in the Bulletin of the [Calcutta Mathematical Society](#) while the former part, despite being submitted for publication in 1941, appeared almost five years later in [Sankhya](#) relative closeness of the two samples being considered.
- It is used to measure the separability of classes in [classification](#) and it is considered to be more reliable than the [Mahalanobis distance](#), as the Mahalanobis distance is a particular case of the Bhattacharyya distance when the standard deviations of the two classes are the same.
- Consequently, when two classes have similar means but different standard deviations, the Mahalanobis distance would tend to zero, whereas the Bhattacharyya distance grows depending on the difference between the standard deviations.

Applications of Bhattacharya Distance

- The coefficient can be used to determine the relative closeness of the two samples being considered.
- It is used to measure the separability of classes in [classification](#) and it is considered to be more reliable than the [Mahalanobis distance](#), as the Mahalanobis distance is a particular case of the Bhattacharyya distance when the standard deviations of the two classes are the same.
- Consequently, when two classes have similar means but different standard deviations, the Mahalanobis distance would tend to zero, whereas the Bhattacharyya distance grows depending on the difference between the standard deviations
- The Bhattacharyya distance is widely used in research of feature extraction and selection, image processing, [speaker recognition](#), and phone clustering.
- A "Bhattacharyya space" has been proposed as a feature selection technique that can be applied to texture segmentation.

Definition of Bhattacharyya distance

- For probability distributions p and q over the same domain X, the Bhattacharyya distance is defined as

$$D_B(p, q) = -\ln(BC(p, q))$$

where

$$BC(p, q) = \sum_{x \in X} \sqrt{p(x)q(x)}$$

- is the Bhattacharyya coefficient for discrete probability distributions.

Bhattacharyya Distance

- **Bhattacharyya distance** measures the similarity of two probability distributions.
- It is closely related to the **Bhattacharyya coefficient** which is a measure of the amount of overlap between two statistical samples or populations.
- Both measures are named after Anil Kumar Bhattacharya, a statistician who worked in the 1930s at the Indian Statistical Institute
- For discrete probability distributions p and q over the same domain X , it is defined as:

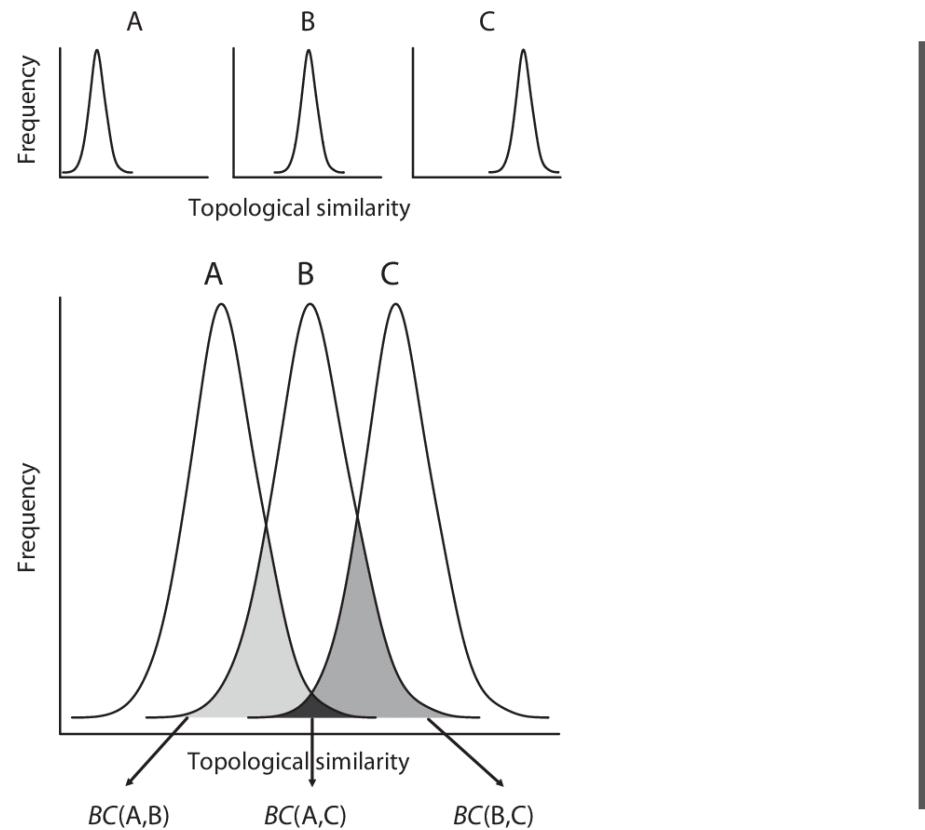
$$D_B(p, q) = -\ln(BC(p, q))$$

- where:
$$BC(p, q) = \sum_{x \in X} \sqrt{p(x)q(x)}$$
- is the Bhattacharyya coefficient. **Bhattacharyya coefficient** is an approximate measurement of the amount of overlap between two statistical samples. The coefficient can be used to determine the relative closeness of the two samples being considered. For continuous distributions, the Bhattacharyya coefficient is defined as:

$$BC(p, q) = \int \sqrt{p(x)q(x)} dx$$

Bhattacharyya coefficient

The Bhattacharyya Coefficient (BC) is the overlap of the distribution of three similarity metrics



Bhattacharyya distance in the Multivariate Normal Case

$$b = \frac{1}{8}(\mu_2 - \mu_1)^T \left[\frac{\Sigma_1 + \Sigma_2}{2} \right]^{-1} (\mu_2 - \mu_1) \\ + \frac{1}{2} \ln \frac{|(\Sigma_1 + \Sigma_2)/2|}{|\Sigma_1|^{1/2} |\Sigma_2|^{1/2}},$$

- where μ_i and Σ_i refer to mean and covariance of ith cluster. $\Sigma = \Sigma_1 + \Sigma_2 / 2$

Things to be taken care for numeric attribute distance calculation

- In some cases data are normalized before applying distance calculations.
- Transforming data to fall within a smaller or common range such as [-1,1] or [0.0,1.0]
- E.g values of height attributes
- Smaller then unit of measurement larger will be the range of values (weight effect)
- Normalization attempts to assign equal weight.

Lecture: 17 Similarity measures for symmetric and asymmetric binary data

22/02/2022

Proximity Measure for binary attribute

- **Nominal attribute** with only 2 states (0 and 1) are known as **Binary attributes**.
- Treating binary variables as if they are interval-scaled can lead to misleading clustering results. Therefore, methods specific to binary data are necessary for computing dissimilarities.
- There are **2 types** of binary attributes
 1. **Symmetric Binary:** A binary attribute that has only 2 outcomes and both outcomes equally important Eg: Male and Female
 2. **Asymmetric Binary:** outcomes of the states not equally important. E.g., medical test (positive vs. negative) Convention: assign 1 to most important outcome (e.g., HIV positive)
- Similarity that is based on symmetric binary variables is called **invariant similarity**.

Contingency Table

- Prepare contingency table for objects i and j, If aboth binary variables are thought of as having the same weight,where
 - q is the number of attributes that equal 1 for both objects i and j
 - r is the number of attributes that equal 1 for object i but that are 0 for object j
 - s is the number of attributes that equal 0 for object i but equal 1 for object j,
 - t is the number of attributes that equal 0 for both objects i and j.
 - p is the total number of attributes where $p = q + r + s + t$.

		Object j		
		1	0	
Object i	1	$q=f_{11}$	$r=f_{10}$	
	0	$s=f_{01}$	$t=f_{00}$	
	Sum	$q+s$	$r+t$	$p=q+r+s+t$

Symmetric Binary Dissimilarity

- Dissimilarity that is based on symmetric binary variables is called symmetric binary dissimilarity.
- Distance measure for symmetric binary variables ie, the dissimilarity between objects i and j:

$$d(i, j) = \frac{r + s}{q + r + s + t}$$

$$= (f_{10} + f_{01}) / (f_{11} + f_{10} + f_{01} + f_{00})$$

Asymmetric Binary Dissimilarity

- Given two asymmetric binary variables, the agreement of two 1s (a positive match) is then considered more significant than that of two 0s (a negative match). Therefore, such binary variables are often considered “monary” (as if having one state).
- The dissimilarity based on such variables is called asymmetric binary dissimilarity. In asymmetric binary dissimilarity the number of negative matches, t , is considered unimportant and thus is ignored in the computation.

$$d(i, j) = \frac{r + s}{q + r + s} = (f_{10} + f_{01}) / (f_{11} + f_{10} + f_{01})$$

EXAMPLE 1

- Suppose that a patient record table contains the **attributes** name, gender, fever, cough, test-1, test2, test-3, and test-4, where name is an **object identifier**, gender is a **symmetric attribute**, and the remaining attributes are **asymmetric binary**. For asymmetric attribute values, let the values Y (yes) and P (positive) be set to 1, and the value N (no or negative) be set to 0. Estimate who all can have similar disease based on dissimilarity measure.

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

EXAMPLE1 SOLUTION

$$d(jack, mary) = \frac{0+1}{2+0+1} = 0.33$$

$$d(jack, jim) = \frac{1+1}{1+1+1} = 0.67$$

$$d(jim, mary) = \frac{1+2}{1+1+2} = 0.75$$

		Mary			
		1	0		Σ_{row}
Jack		1	2	0	2
0		1	3	4	
Σ_{co}	1	3	3	6	

		Jim			
		1	0		Σ_{row}
Jack		1	1	1	2
0		1	3	4	
Σ_{co}	1	2	4	6	

		Mary			
		1	0		Σ_{row}
Jim		1	1	1	2
0		2	2	4	
Σ_{col}	1	3	3	6	

- The distance between each pair of the three patients, Jack, Mary, and Jim, is :
- These measurements suggest that Jim and Mary are unlikely to have a similar disease because they have the highest dissimilarity value among the three pairs. Of the three patients, Jack and Mary are the most likely to have a similar disease

How to calculate proximity measure for asymmetric binary attributes?

- Contingency table for binary data

Object 2				
Object 1		1 / True / Positive	0 / False / Negative	Sum
	1 / True / Positive	A	B	A + B
	0 / False / Negative	C	D	C + D
	Sum	A + C	B + D	

Facts observed from table

- A represents that object 1 is True and object 2 is also True.
- B represents that object 1 is True and object 2 is also False.
- C represents that object 1 is False and object 2 is also True.
- D represents that object 1 is False and object 2 is also False.

Object 2				
Object 1		1 / True / Positive	0 / False / Negative	Sum
	1 / True / Positive	A	B	A + B
	0 / False / Negative	C	D	C + D
	Sum	A + C	B + D	

Additional Examples

Name	Fever	Cough	Test 1	Test 2	Test 3	Test 4
CRIS	Negative	Yes	Negative	Positive	Negative	Negative
RAM	Negative	Yes	Negative	Positive	Positive	Negative
SHAM	Positive	Yes	Negative	Negative	Negative	Negative

Formula

$$= \frac{B + C}{A + B + C} = 0.67$$

Solution

- Distance (object1, Object2) = $B+C / A+B+C$
- Distance(Chris, Sham) = $1+1 / 1+1+1 = 2/3 = 0.67$
- Distance(Chris, Ram) = $0+1 / 2+0+1 = 1/3 = 0.33$
- Distance(Ram, Sham) = $1+2 / 1+1+2 = 3/4 = 0.75$

Remark 1: Asymmetric Data

- In table 2, Chris, Ram and Sham are objects.
- Negative values represents False and Positive represents Negative.
- dissimilarity of binary variables
- In the results, we can see the following facts;
- The distance between object 1 and 2 is 0.67. Chris is object 1 and Sham is in object 2 and the distance between both is 0.67.
- Less distance is between Chris and Ram. It means that Chris and Ram are more similar to each other as compared to other objects.

Distance measure for symmetric binary variables

- **How to calculate proximity measure for symmetric binary attributes?**

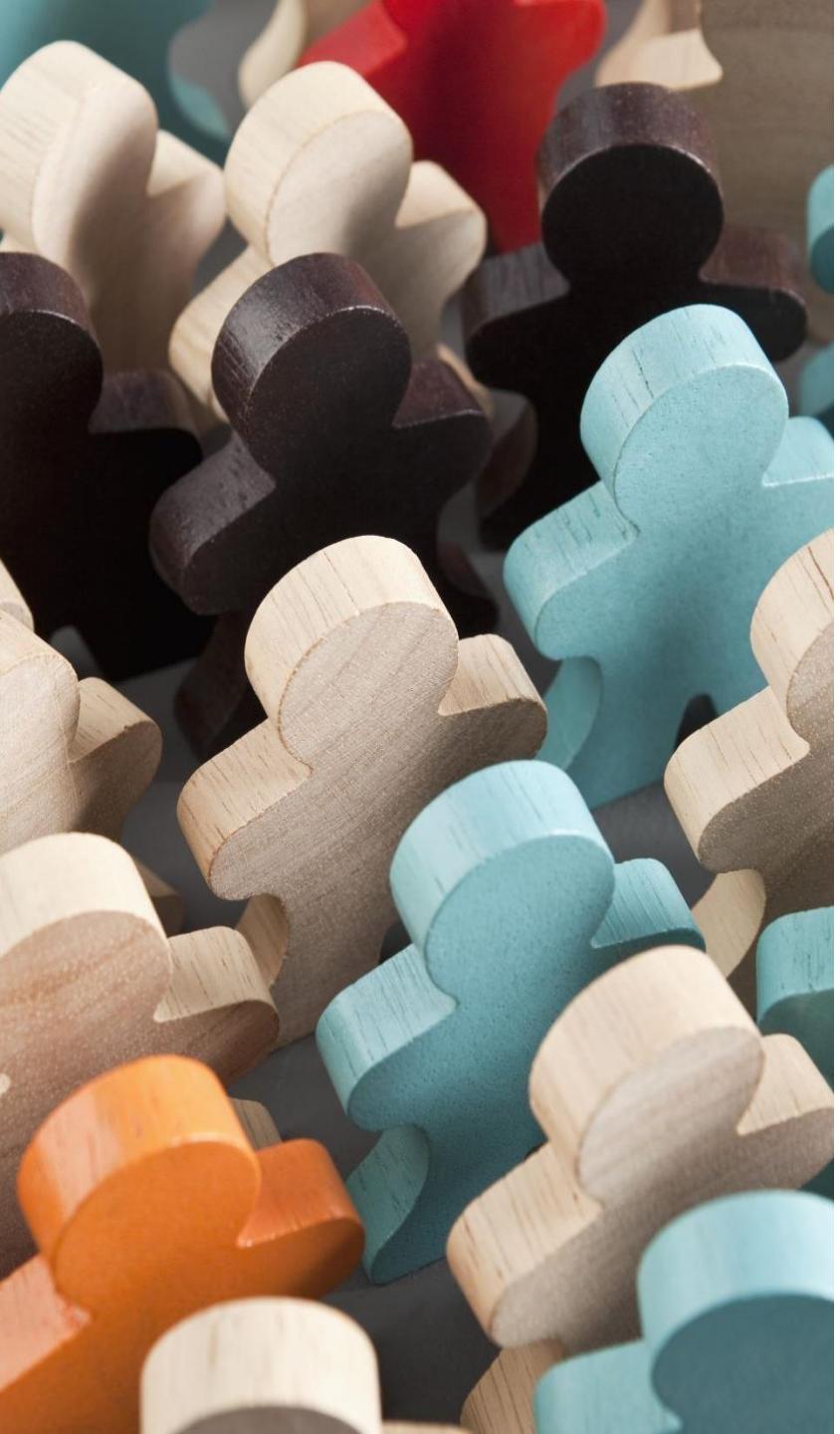
Another Example-2

Object 2				
Object 1		1 / True / Positive	0 / False / Negative	Sum
	1 / True / Positive	A	B	A + B
	0 / False / Negative	C	D	C + D
	Sum	A + C	B + D	

Name	Gender	Job_Status
Chirag	Male	Regular
Rajesh	Male	Contract

Remark 2: Symmetric Data

- Consider 1 for positive/True and 0 for negative/False
- Here we are considering Male and regular as positive and female and contract as negative.
- **A** = Chirag is positive and Rajesh is also positive. so A=1 because Rajesh and Chirag both are male and the male is positive.
- **B** = Chirag is positive and Rajesh is negative. So B=1 because Chirag is regular that is positive and Rajesh is on contract that is negative
- **C** = Chirag is negative and Rajesh is 1. So C = 0 because Chirag is never negative. He is male and regular. and male and regular both are positive.
- **D** = Chirag is negative and Rajesh is also negative. So D=0 because Chirag is never negative. He is always positive(male and regular).



Proximity Measure of Nominal Attributes/ Categorical Attributes

- The values of a Nominal attribute are categories, states, or “names of things”. It is referred as categorical attributes or enumerations. The values do not have any meaningful order(rank, position) . A **nominal attribute** can take **2 or more states**.

- Example: Color (red, yellow, blue, green), profession.

“How is dissimilarity computed between objects described by nominal attributes?”

- **Simple Matching Method:** The **dissimilarity between two objects i and j** can be computed based on the ratio of mismatches:

$d(i,j)=(p-m)/p$ Where **m**: #number of matches(i.e., the number of attributes for which i and j are in the same state), **p**: total #number of variables/attributes **describing the objects**

- **similarity** can be computed as

$$sim(i, j) = 1 - d(i, j) = \frac{m}{p}.$$

EXAMPLE

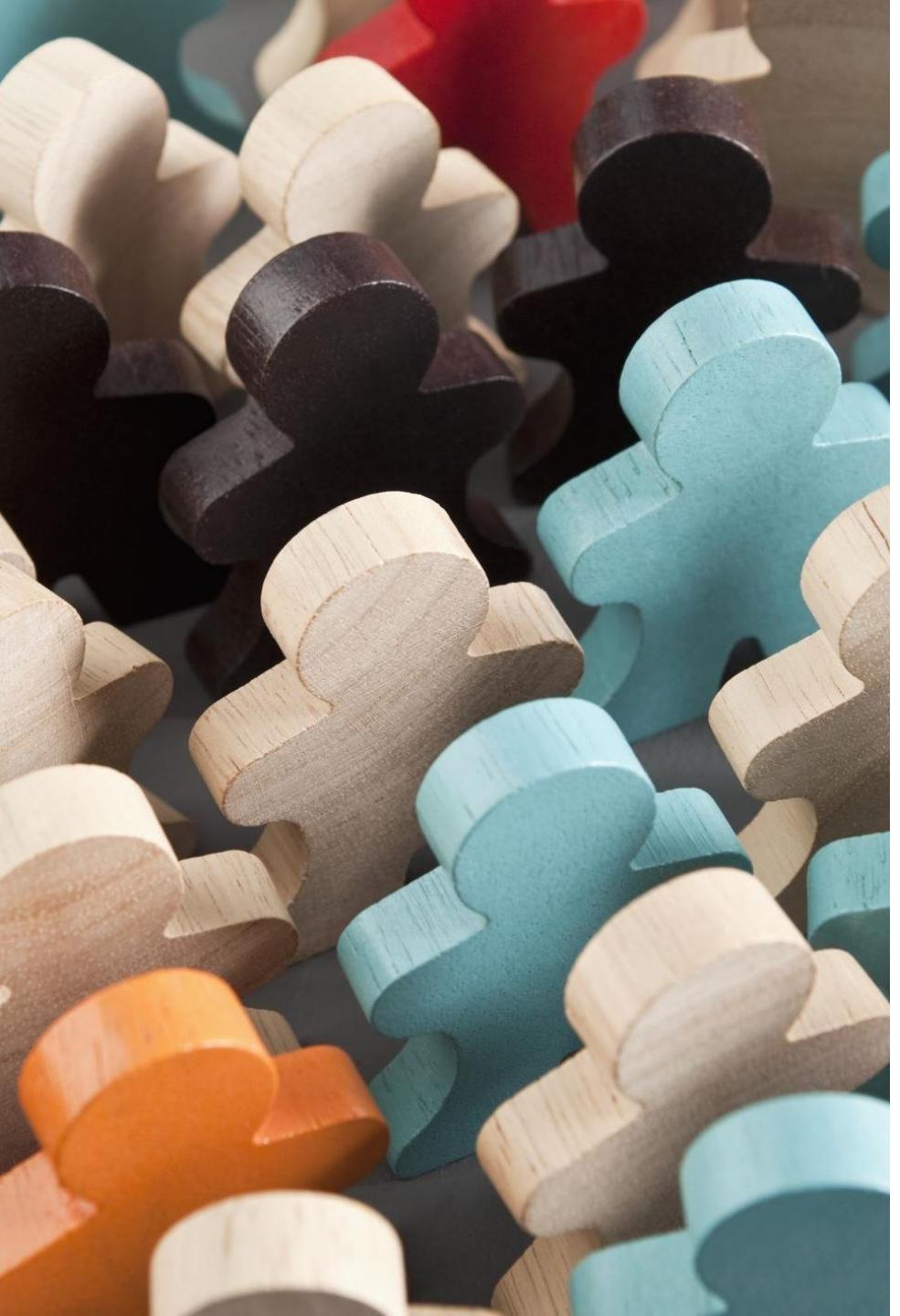
- Consider the nominal data given here.

<i>object identifier</i>	<i>test-1</i> (nominal)	<i>test-2</i> (ordinal)	<i>test-3</i> (numeric)
1	code-A	excellent	45
2	code-B	fair	22
3	code-C	good	64
4	code-A	excellent	28

$$\begin{bmatrix} 0 \\ d(2, 1) & 0 \\ d(3, 1) & d(3, 2) & 0 \\ d(4, 1) & d(4, 2) & d(4, 3) & 0 \end{bmatrix}$$

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

here we have one nominal attribute, test-1, we set $p = 1$ so that $d(i, j)$ evaluates to 0 if objects i and j match, and 1 if the objects differ.



Proximity Measure of Nominal Attributes: Alternative approach

- Nominal attributes can be encoded by asymmetric binary attributes by creating a new binary attribute for each of the M states. For an object with a given state value, the binary attribute representing that state is set to 1, while the remaining binary attributes are set to 0.
- **Example:** To encode the nominal attribute **map color**, a binary attribute can be created for each of the **five colors** red, yellow, blue, green and white. For an object having the color yellow, the yellow attribute is set to 1, while the remaining four attributes are set to 0.
Example: Nominal attribute *educationlevel*: has three values,
 - *Primary school, High school, university*
 - We create three binary attributes.
 - If a particular data instance in the original data has *university* as the value for *educationlevel*,
 - then in the transformed data, we set the value of the attribute *university* to 1, and
 - the values of attributes *primaryschool* and *highschool* to 0



Proximity Measures for Ordinal Attributes

- The values of an ordinal attribute have a meaningful order or ranking about them, yet the magnitude between successive values is unknown.
- An ordinal variable can be discrete or continuous.
 - Replace an ordinal variable value by its rank: $r_{if} \in \{1, \dots, M_f\}$
 - Normalize the rank by mapping the range of each variable onto [0, 1] by replacing i -th object in the f -th variable by using
$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$
 - Dissimilarity can then be computed using any of the distance measures described prior for numeric attributes, using z_{if} to represent the f value for the i th object
- The value of f for the i th object is x_{if} , and f has M_f ordered states, representing the ranking $1, \dots, M_f$. Replace each x_{if} by its corresponding rank, r_{if} $\{1, \dots, M_f\}$.

EXAMPLE

- Consider the nominal data given in test-2.

<i>object identifier</i>	<i>test-1 (nominal)</i>	<i>test-2 (ordinal)</i>	<i>test-3 (numeric)</i>
1	code-A	excellent	45
2	code-B	fair	22
3	code-C	good	64
4	code-A	excellent	28

There are three states for test -2, namely **fair**, **good**, and **excellent**, that is $M_f = 3$.

- Replace each value for **test -2 by its rank**, the four objects are assigned the **ranks 3, 1, 2, and 3**, respectively.
- Normalizes the ranking by mapping **rank 1** to **0.0[1-1/3-1]**, **rank 2** to **0.5[2-1/3-1]**, and **rank 3** to **1.0[3-1/3-1]**.
- Use Euclidean distance for dissimilarity calculation

$$\begin{bmatrix} 0 \\ 1.0 & 0 \\ 0.5 & 0.5 & 0 \\ 0 & 1.0 & 0.5 & 0 \end{bmatrix}$$

EXAMPLE

- Given an ordinal data: freshman, sophomore, junior, senior.
Compute dissimilarity matrix using Manhattan Distance.

SOLUTION

1. Replace an ordinal variable value by its rank: freshman: 0; sophomore: 1; junior: 2; senior 3
2. Normalize rank: freshman: $0((1-1)/(4-1))$; sophomore: $1/3((2-1)/(4-1))$; junior: $2/3((3-1)/(4-1))$; senior $1((4-1)/(4-1))$
3. Dissimilarity matrix using Manhattan Distance

0			
$1/3$	0		
$2/3$	$1/3$	0	
1	$2/3$	$1/3$	0

Similarity and dissimilarity measure for single attribute

- The following table shows the similarity and dissimilarity between two objects, x and y , with respect to a single, simple attribute.

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$	$s = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases}$
Ordinal	$d = x - y /(n - 1)$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - d$
Interval or Ratio	$d = x - y $	$s = -d, s = \frac{1}{1+d}, s = e^{-d}, s = 1 - \frac{d - \min_d}{\max_d - \min_d}$

Lecture: 18 LCS and Cosine Distance

24/02/2022

Cosine Similarity

- In data analysis, cosine similarity is a measure of similarity between two sequences of numbers.
- For defining it, the sequences are viewed as vectors in an inner product space, and the cosine similarity is defined as the cosine of the angle between them, that is, the dot product of the vectors divided by the product of their lengths.
- It follows that the cosine similarity does not depend on the magnitudes of the vectors, but only on their angle.
- The cosine similarity always belongs to the interval [-1,1].
- For example, two proportional vectors have a cosine similarity of 1, two orthogonal vectors have a similarity of 0, and two opposite vectors have a similarity of -1.
- The cosine similarity is particularly used in positive space, where the outcome is neatly bounded in [0,1].

Cosine Similarity Formula

$$\text{similarity}(A, B) = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n A_i^2} \times \sqrt{\sum_{i=1}^n B_i^2}}$$

Computation of Cosine Similarity

- Given the following two vectors compute the cosine similarity
- D1= 4 0 2 0 1
- D2= 2 0 0 2 2
- Solution:
- $D1 \bullet D2 = 4*2 + 0*0 + 2*0 + 0*2 + 1*2 = 10$
- $||D1|| = (4^2 + 2^2 + 1^2)^{0.5} = (16+4+1)^{0.5} = (21)^{0.5} = 4.58$
- $||D2|| = (2^2 + 2^2 + 2^2)^{0.5} = (4+4+4)^{0.5} = (12)^{0.5} = 3.46$
- $\text{COS}(D1, D2) = (D1 \bullet D2) / (||D1|| * ||D2||)$
 $= 10 / (4.58 * 3.46)$
 $= 0.63$

LCS Problem

- The longest common subsequence (LCS) problem is the problem of finding the longest subsequence common to all sequences in a set of sequences (often just two sequences).
- It differs from the longest common substring problem: unlike substrings, subsequences are not required to occupy consecutive positions within the original sequences.
- The longest common subsequence problem is a classic computer science problem, the basis of data comparison programs such as the diff utility, and has applications in computational linguistics and bioinformatics.
- It is also widely used by revision control systems such as Git for reconciling multiple changes made to a revision-controlled collection of files.

Example

- consider the sequences (ABCD) and (ACBAD).
- They have 5 length-2 common subsequences:
 - (AB), (AC), (AD), (BD), and (CD);
- 2 length-3 common subsequences: (ABD) and (ACD); and no longer common subsequences.
- So (ABD) and (ACD) are their longest common subsequences

LCS

- The LCS problem has an optimal substructure: the problem can be broken down into smaller, simpler subproblems, which can, in turn, be broken down into simpler subproblems, and so on,
- until, finally, the solution becomes trivial.
- LCS in particular has overlapping subproblems: the solutions to high-level subproblems often reuse solutions to lower level subproblems.
- Problems with these two properties are amenable to dynamic programming approaches, in which subproblem solutions are memoized, that is, the solutions of subproblems are saved for reuse.

DP Algorithm

LCS-Length(X, Y)

1. $m = \text{length}(X)$ // get the # of symbols in X
2. $n = \text{length}(Y)$ // get the # of symbols in Y
3. for $i = 1$ to m $c[i,0] = 0$ // special case: $Y[0]$
4. for $j = 1$ to n $c[0,j] = 0$ // special case: $X[0]$
5. for $i = 1$ to m // for all $X[i]$
6. for $j = 1$ to n // for all $Y[j]$
7. if ($X[i] == Y[j]$)
8. $c[i,j] = c[i-1,j-1] + 1$
9. else $c[i,j] = \max(c[i-1,j], c[i,j-1])$
10. return c



Longest Common Subsequence Algorithm

- A subsequence of a string S , is a set of characters that appear in left-to-right order, but not necessarily consecutively.
- **Example:** Consider ACTTGCG
 - ACT , ATTC , T , ACTTGC are all subsequences. TTA is not a subsequence. There are 2^n subsequences of string S of length n .
- A common subsequence of two strings is a subsequence that appears in both strings. A longest common subsequence is a common subsequence of maximal length.

Example

$S_1 = \text{AAACCGTGAGTTATTCGTTCTAGAA}$

$S_2 = \text{CACCCTAAGGTACCTTTGGTTC}$

LCS is ACCTAGTACTTG

RECURSIVE ALGORITHM

- If last characters of both sequences match (or $X[m-1] == Y[n-1]$) then $L(X[0..m-1], Y[0..n-1]) = 1 + L(X[0..m-2], Y[0..n-2])$
- If last characters of both sequences do not match (or $X[m-1] != Y[n-1]$) then $L(X[0..m-1], Y[0..n-1]) = \text{MAX} (L(X[0..m-2], Y[0..n-1]), L(X[0..m-1], Y[0..n-2]))$

Examples:

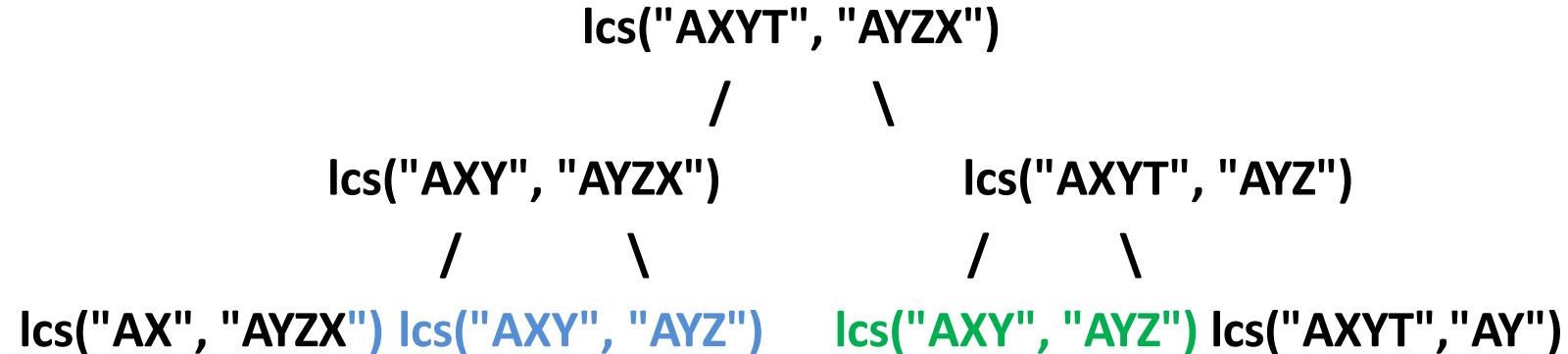
- 1) Consider the input strings “**AGGTAB**” and “**GXTXAYB**”. Last characters match for the strings. So length of LCS can be written as: $L(\text{“AGGTAB”}, \text{“GXTXAYB”}) = 1 + L(\text{“AGGTA”}, \text{“GXTXAY”})$
- 2) Consider the input strings “**ABCDGH**” and “**AEDFHR**”. Last characters do not match for the strings. So length of LCS can be written as: $L(\text{“ABCDGH”}, \text{“AEDFHR”}) = \text{MAX} (L(\text{“ABCDG”}, \text{“AEDFHR”}), L(\text{“ABCDGH”}, \text{“AEDFH”}))$

RECURSIVE ALGORITHM

```
LCS(i,j){  
    If(A[i] ≠ null || B[j]≠null)  
        return 0;  
    Elseif(A[i]==B[j])  
        return (1+ LCS(i+1,j+1))  
    Else  
        return(max(LCS(i+1,j),LCS(i,j+1)))  
}
```

Overlapping Subproblem

- This is a correct solution but it's very time consuming. For example, if the two strings have no matching characters, so the last line always gets executed, the time bounds are binomial coefficients, which (if $m=n$) are close to 2^n .



- In the above partial recursion tree, `lcs("AXY", "AYZ")` is being solved twice. So this problem has Overlapping Substructure property and re-computation of same sub problems can be avoided by either using Memoization or Tabulation by using "top down" approach of dynamic programming. The concept is to cache the result of a function given its parameter so that the calculation will not be repeated; it is simply retrieved, or memo-ed. Most of the time a simple array is used for the cache table, but a hash table or map could also be employed.

DYNAMIC PROGRAMMING APPROACH

Theorem: Let $X = \langle x_1, x_2, \dots, x_m \rangle$ and $Y = \langle y_1, y_2, \dots, y_n \rangle$ be sequences, and let $Z = \langle z_1, z_2, \dots, z_k \rangle$ be any LCS of X and Y .

1. If $x_m = y_n$, then $z_k = x_m = y_n$ and Z_{k-1} is an LCS of X_{m-1} and Y_{n-1} .
 2. If $x_m \neq y_n$, then $z_k \neq x_m$ implies that Z is an LCS of X_{m-1} and Y_n .
 3. If $x_m \neq y_n$, then $z_k \neq y_n$ implies that Z is an LCS of X and Y_{n-1} .
- So dynamic way of solving LCS is

$$LCS[i, j] = \begin{cases} \text{if}(A[i]==B[j]) \\ \quad \quad \quad \text{return } (1 + LCS(i-1, j-1)) \\ \text{Else} \\ \quad \quad \quad \text{return } \max(LCS[i, j - 1], LCS[i - 1, j]) \end{cases}$$

ALGORITHM

```
LCS – Length(X, Y )  
m= length[X]  
n= length[Y ]  
for i =1 to m  
    c[i, 0] = 0  
for j =0 to n  
    c[0, j] = 0  
for i = 1 to m  
    for j = 1 to n  
        if xi == yj  
            c[i, j] = c[i - 1, j - 1] + 1  
            B[i, j] := 'D' or ↗  
        else if (c[i - 1, j] ≥ c[i, j - 1])  
            c[i, j] = c[i - 1, j]  
            B[i, j] := 'U' or ↑  
        else  
            c[i, j] = c[i, j - 1]  
            B[i, j] := 'L' or ←  
return c and B
```

SEQUENCE RETRIEVAL

Algorithm: Print-LCS (B, X, i, j)

if $i = 0$ and $j = 0$
return

if $B[i, j] = 'D'$ or ↙
 Print-LCS(B, X, i-1, j-1)
 Print(x_i)

else if $B[i, j] = 'U'$ or ↑
 Print-LCS(B, X, i-1, j)

else
 Print-LCS(B, X, i, j-1) //for 'L' or ←

EXAMPLE 1

- we have two strings $X = \mathbf{ABCBDAB}$ and $Y = \mathbf{BDCABA}$ to find the longest common subsequence.
Following the algorithm LCS-Length-Table-Formulation

j	0	1	2	3	4	5	6
i	y_j	B	D	C	A	B	A
0	x_i	0	0	0	0	0	0
1	A	0	0	0	0	1	1
2	B	0	1	-1	-1	1	-2
3	C	0	1	1	2	-2	2
4	B	0	1	1	2	2	-3
5	D	0	1	2	2	3	3
6	A	0	1	2	2	3	4
7	B	0	1	2	2	3	4

EXAMPLE 2

X= innovation and Y= tionwagon LCS=inaon

	Yj	T	I	O	N	W	A	G	O	N
Xi	0	0	0	0	0	0	0	0	0	0
I	0	0	1 ↑	1 ↘	1 ←	1 ←	1 ←	1 ←	1 ←	1 ←
N	0	0	1 ↑	1 ↑	2 ↘	2 ←	2 ↑	2 ←	2 ←	2 ↘
N	0	0	1	1 ↑	2 ↘	2 ↑	2 ↑	2 ↑	2 ↑	3 ↘
O	0	0	1	2 ↘	2	2	2	2	3 ↘	3
V	0	0	1	2	2	2	2	2	3	3
A	0	0	1	2	2	2	3 ↘	3 ←	3	3
T	0	1 ↘	1	2	2	2	3	3	3	3
I	0	1	2 ↘	2	2	2	3	3	3	3
O	0	1	2	3 ↘	3 ←	3 ←	3	3	4 ↘	4
N	0	1	2	3 ↑	4 ↘	4 ←	4 ←	4 ←	4 ←	5 ↘

Lecture: 19 Textual Similarity Measure

28/02/2022



Similarity measures for symmetric and asymmetric binary data

1. Simple matching coefficient
2. Jaccard coefficient
3. Hamming distance
4. Jaro Distance

Agenda

- simple matching coefficient, Jaccard coefficient, hamming distance
- Jaro distance,
- Similarity measures for textual data,

Simple Matching coefficient

What is Simple Matching Coefficient(SMC)

- The **simple matching coefficient (SMC)** or **Rand similarity coefficient** is a statistic used for comparing the similarity and diversity of sample sets.

Definition SMC

- Given two objects, A and B, each with n binary attributes, SMC is defined as:

$$\begin{aligned} \text{SMC} &= \frac{\text{number of matching attributes}}{\text{number of attributes}} \\ &= \frac{M_{00} + M_{11}}{M_{00} + M_{11} + M_{01} + M_{10}} \end{aligned}$$

where:

M_{00} is the total number of attributes where A and B both have a value of 0.

M_{11} is the total number of attributes where A and B both have a value of 1.

M_{01} is the total number of attributes where the attribute of A is 0 and the attribute of B is 1.

M_{10} is the total number of attributes where the attribute of A is 1 and the attribute of B is 0.

		A	
		0	1
B	0	M_{00}	M_{10}
	1	M_{01}	M_{11}

Simple Matching Distance (SMD)

- The simple matching distance (SMD), which measures dissimilarity between sample sets, is given by : $1 - SMC$
- SMC is linearly related to Hamann similarity:

$$SMC = (Hamann + 1)/2. \text{Also, } SMC = 1 - D^2/n,$$

where D^2 is the squared Euclidean distance between the two objects (binary vectors) and n is the number of attributes.

Example : Let assume Two set of Binary

a	0	1	1	0	1
b	1	1	1	0	0

Solution SMC:

- Find the $M_{00} = a=0 b=0 \rightarrow 1$
 - Find the $M_{01} = a=0 b=1 \rightarrow 1$
 - Find the $M_{10} = a=1 b=0 \rightarrow 1$
 - Find the $M_{11} = a=1 b=1 \rightarrow 2$
-
- Fit the value in Formula: $SMC = (M_{11} + M_{00}) / (M_{00} + M_{01} + M_{10} + M_{11})$
 $= 2+1/2+1+1+1=3/5=0.6$

Symmetric Binary similarity: Simple matching coefficient

- **Simple matching coefficient** is a similarity coefficient defined for symmetric binary attributes. It is defined as:

$$S = 1 - D = 1 - \frac{r + s}{q + r + s + t}$$

$$= q+t/q+r+s+t$$

$= (f_{11} + f_{00}) / (f_{01} + f_{10} + f_{11} + f_{00})$ = number of matches / number of attributes

- This measure counts both presences and absences equally.
- **Application:** To find students who had answered similarly in a test that consist of only true or false answers.

Asymmetric Binary Similarity: **Jaccard coefficient**

- Complementarily, we can measure the difference between two binary attributes based on the notion of similarity instead of dissimilarity.
- The asymmetric binary similarity between the objects i and j can be computed as,

$$\text{sim}(i, j) = \frac{q}{q+r+s} = 1 - d(i, j)$$

= number of 11 matches / number of non-zero attributes

$$= (f_{11}) / (f_{01} + f_{10} + f_{11})$$

- The coefficient $\text{sim}(i, j)$ is called the **Jaccard coefficient**

EXAMPLE 1

- **Given:** $x = 100000000$ and $y = 000001001$.
Calculate SMC and Jaccard's coefficient.

SOLUTION

- $f_{01} = 2$ (the number of attributes where **x** was 0 and **y** was 1)
- $f_{10} = 1$ (the number of attributes where **x** was 1 and **y** was 0)
- $f_{00} = 7$ (the number of attributes where **x** was 0 and **y** was 0)
- $f_{11} = 0$ (the number of attributes where **x** was 1 and **y** was 1)
- SMC = $(f_{11} + f_{00}) / (f_{01} + f_{10} + f_{11} + f_{00}) = (0+7) / (2+1+0+7) = 0.7$
- J = $(f_{11}) / (f_{01} + f_{10} + f_{11}) = 0 / (2 + 1 + 0) = 0$

EXAMPLE 2

- Consider the following dataset, where objects are defined with binary attributes. Gender = {M, F}, Hobby = {T, C}, Job = {Y, N} Food = {V, N}, Caste = {H, M}, Education = {L, I} How you can calculate similarity between these 2 people based on similarity measure if Gender, Hobby and Job are symmetric binary attributes and Food, Caste, Education are asymmetric binary attributes?V-1,M-1,L-1

Object	Gender	Food	Caste	Education	Hobby	Job
Hari	M	V	M	L	C	N
Ram	M	N	M	I	T	N
Tomi	F	N	H	L	C	Y

Solution

- Obtain the similarity matrix with Jaccard coefficient of objects

$$\mathcal{J} = \begin{matrix} & H & R & T \\ H & 0 & 0 & 0 \\ R & J(R,H) & 0 & 0 \\ T & J(T,H) & J(T,R) & 0 \end{matrix} \quad \begin{matrix} \bullet J(\text{Hari}, \text{Ram})=1/(1+2+0)=0.33 \\ \bullet J(\text{Ram}, \text{Tomi}) = 0/0+1+1 = 0 \\ \bullet J(\text{Tom},\text{Hari})=1/1+2+0=0.33 \end{matrix}$$

Jaccard coefficient

Jaccard coefficient

- So far discussed some metrics to find the similarity between objects. where the objects are points or vectors. When we consider Jaccard similarity these objects will be sets.

Jaccard Similarity

Set A = {  }

Set B = {  }

$|A| = 4$ $|B| = 5$ [@dataaspirant.com](http://dataaspirant.com)

$$\text{Jaccard Similarity } J(A, B) = | \text{Intersection}(A, B) | / | \text{Union}(A, B) |$$

$$= 2 / 7$$

$$= 0.286$$

Jaccard Coefficient

- The Jaccard similarity index (sometimes called the Jaccard similarity *coefficient*) compares members for two sets to see which members are shared and which are distinct.
- It's a measure of similarity for the two sets of data, with a range from 0% to 100%.
- The higher the percentage, the more similar the two populations.
- Although it's easy to interpret, it is extremely sensitive to small samples sizes and may give erroneous results, especially with very small samples or data sets with missing observations.

Need of Jaccard Coefficient over SMC

- Number of Coefficient is not similar
- Example of shop where customer purchase to items but he forgot list
- If number of zeroes are more than no of ones then SMC not work properly and hence necessary to introduce the Jaccard coefficient

How to Calculate the Jaccard Index/Coefficient?

- Jaccard Index =
 $(\text{the number in both sets}) / (\text{the number in either set}) * 100$

Formula:

$$J(X, Y) = |X \cap Y| / |X \cup Y|$$

Similar SMC;

OR JC Similarity = Number of Positive overlapping/ Total Positive

OR Jaccard can be computed :

$$Jc = M_{11} / (M_{01} + M_{10} + M_{11})$$

STEPS

- 1.Count the number of members which are shared between both sets.
- 2.Count the total number of members in both sets (shared and un-shared).
- 3.Divide the number of shared members (1) by the total number of members (2).
- 4.Multiply the number you found in (3) by 100.

INTERPRETATION - JACCARD SIMILARITY

- **This percentage tells you how similar the two sets are.**
- Two sets that share all members would be 100% similar. the closer to 100%, the more similarity (e.g. 90% is more similar than 89%).
- If they share no members, they are 0% similar.
- The midway point — 50% — means that the two sets share half of the members.

Example 2: Jaccard Coefficient

- Researchers are studying biodiversity in two rainforests.
- They catalog specimens from six different species, A,B,C,D,E,F.
- Two species are shared between the two rainforests.
- What is the Jaccard coefficient?

Solution 2: Jaccard coefficient

1.Two species (3 and 5) are shared between both populations.

2.There are 6 unique species in the two populations.

$$3/6 = 1/3$$

$$4.1/3 * 100 = 33.33\%.$$

- Rainforests A and B are 33% similar.

EXAMPLE: FIND SIMILARITY BETWEEN DISTANCES?

- $A = \{0,1,2,5,6\}$
- $B = \{0,2,3,4,5,7,9\}$
- Solution:
- $J(A,B) = |A \cap B| / |A \cup B| = |\{0,2,5\}| / |\{0,1,2,3,4,5,6,7,9\}| = 3/9 = 0.33.$
- Note:
 - 1.The cardinality of A, denoted $|A|$ is a count of the number of elements in set A.
 - 2.Although it's customary to leave the answer in decimal form if you're using set notation, you could multiply by 100 to get a similarity of 33.33%.

Hamming Distance

Hamming Distance

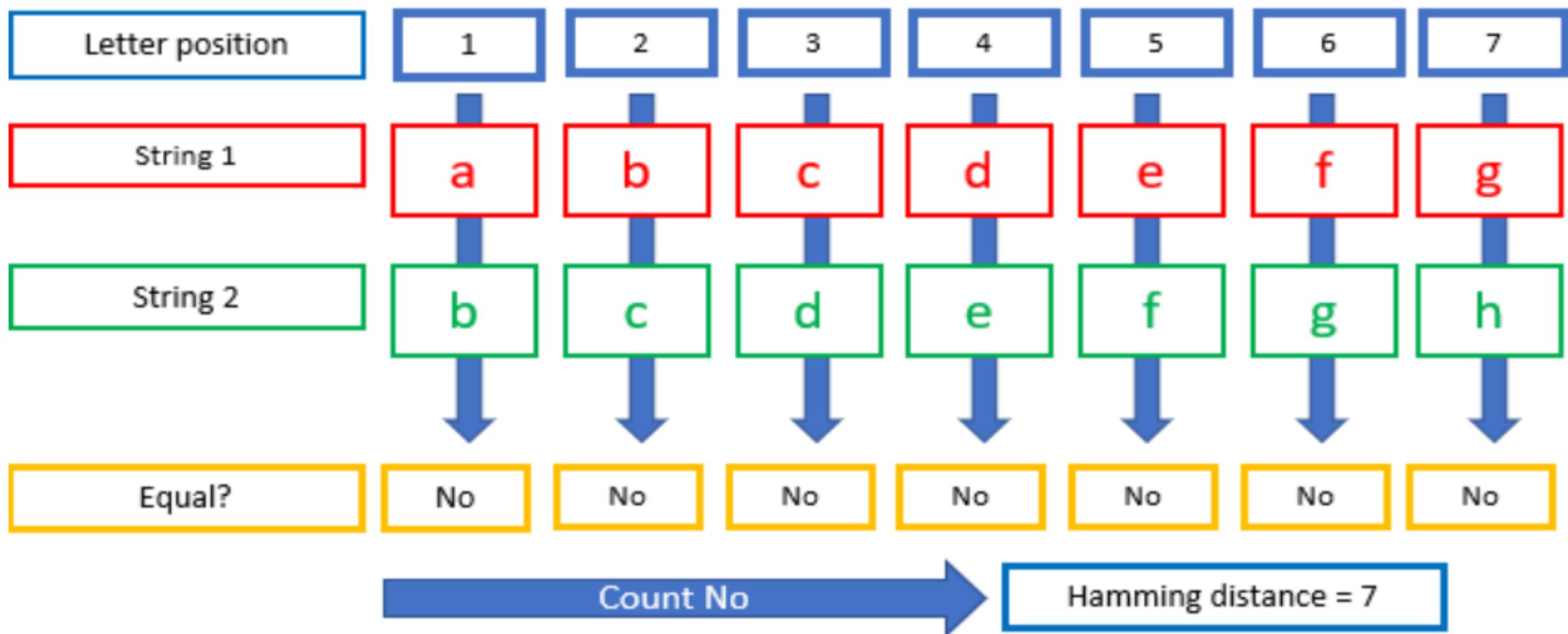
- The Hamming Distance compares every letter of the two strings based on position.
- So the first letter of word 1 is compared to the first letter of word 2 etc .
- The Hamming Distance compares every letter of the two strings based purely on position.
- To compute the Hamming distance between two strings, you compare the characters of each position in the string.
- The number of unequal characters is the Hamming distance.

Additional Example- Hamming distance

- *Hamming Distance measures the distance between two strings of the same length. Hamming distance $d=q+r$.*
- *The Hamming Distance between two strings of the same length is the number of bit positions at which the corresponding characters or bits are different.*
- Example: $x = 010101001$ and $y = 010011000$
- Hamming distance = 3; there are 3 binary numbers different between the x and y.
 $x = 010\textcolor{red}{1}01\textcolor{red}{0}00\textcolor{red}{1}$
 $y = 010\textcolor{red}{0}01\textcolor{red}{1}000$
- Hamming distance can be treated as a special instance of Manhattan distance, when attribute values $\in [0, 1]$ is called Hamming distance.
- Hamming distance is used to measure the distance between categorical variables
- *Hamming Distance=7*

Hamming distance

From: '*abcdefg*' To: '*bcdefgh*'



Opinion on Hamming Distance

- Advantages- Very fast and simple to do this position-wise comparison.
- On the other hand, critics are that it cannot take into account two strings with an unequal number of letters.
- Another critic is that it is too strict, for example, “*abcdefg*” and “*bcdefgh*” are considered totally different, while 6 out of 7 characters are the same.

Similarity measures for textual data

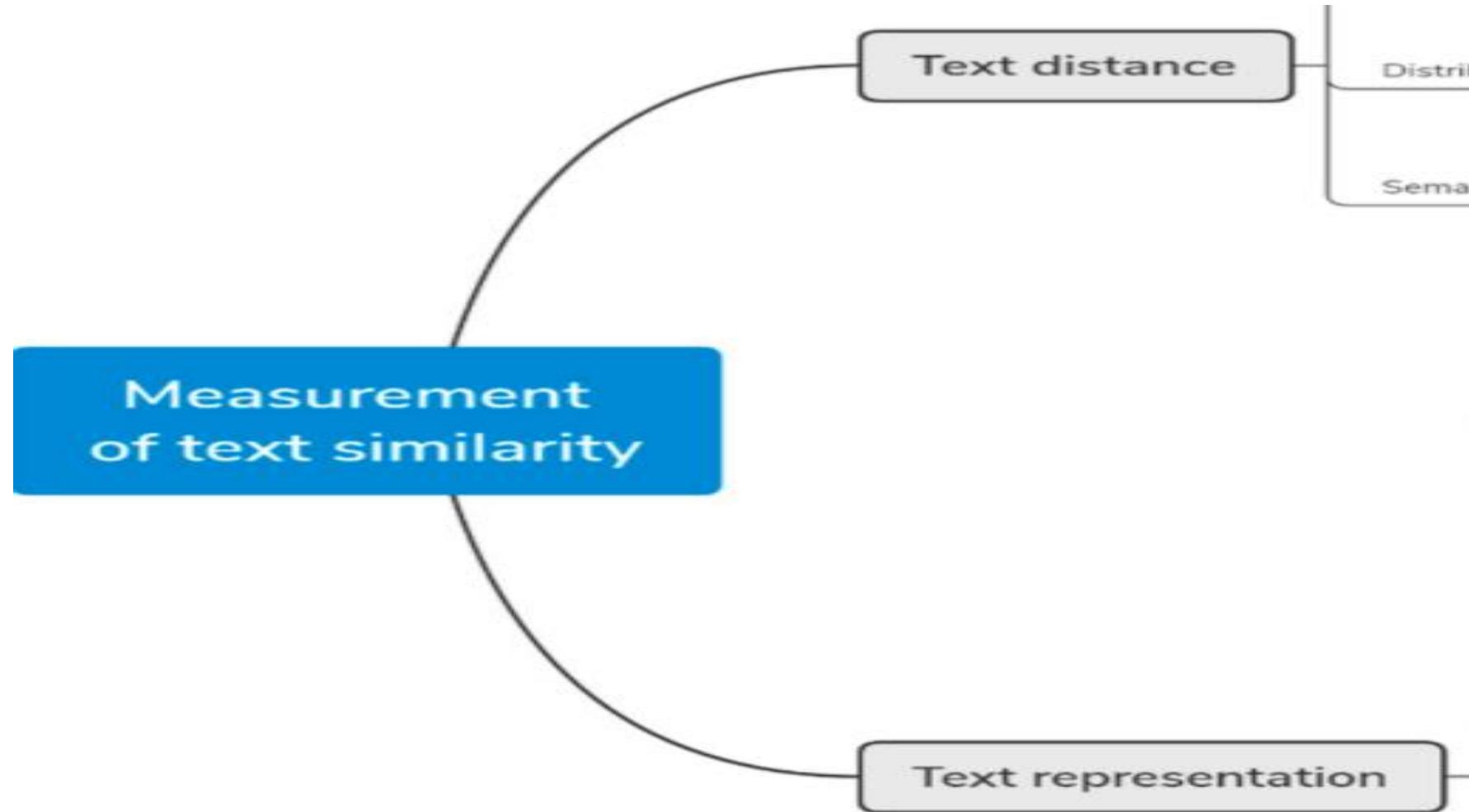
Similarity features of Text

Algorithm	Comparison by	Case Sensitive	Sequence Matters?	Result
Cosine Similarity / Distance	word	✓	X	Number (0 to 1)
Hamming Distance (Same length input strings only)	character	✓	✓	Count of substitutions
Jaccard Similarity / Distance	character	✓	X	Number (0 to 1)
Jaro Winkler Similarity / distance	character	✓	✓	Number (0 to 1)
Levenshtein Distance	character	✓	✓	Count of edits
Longest Common Subsequence	character	✓	✓	Common String

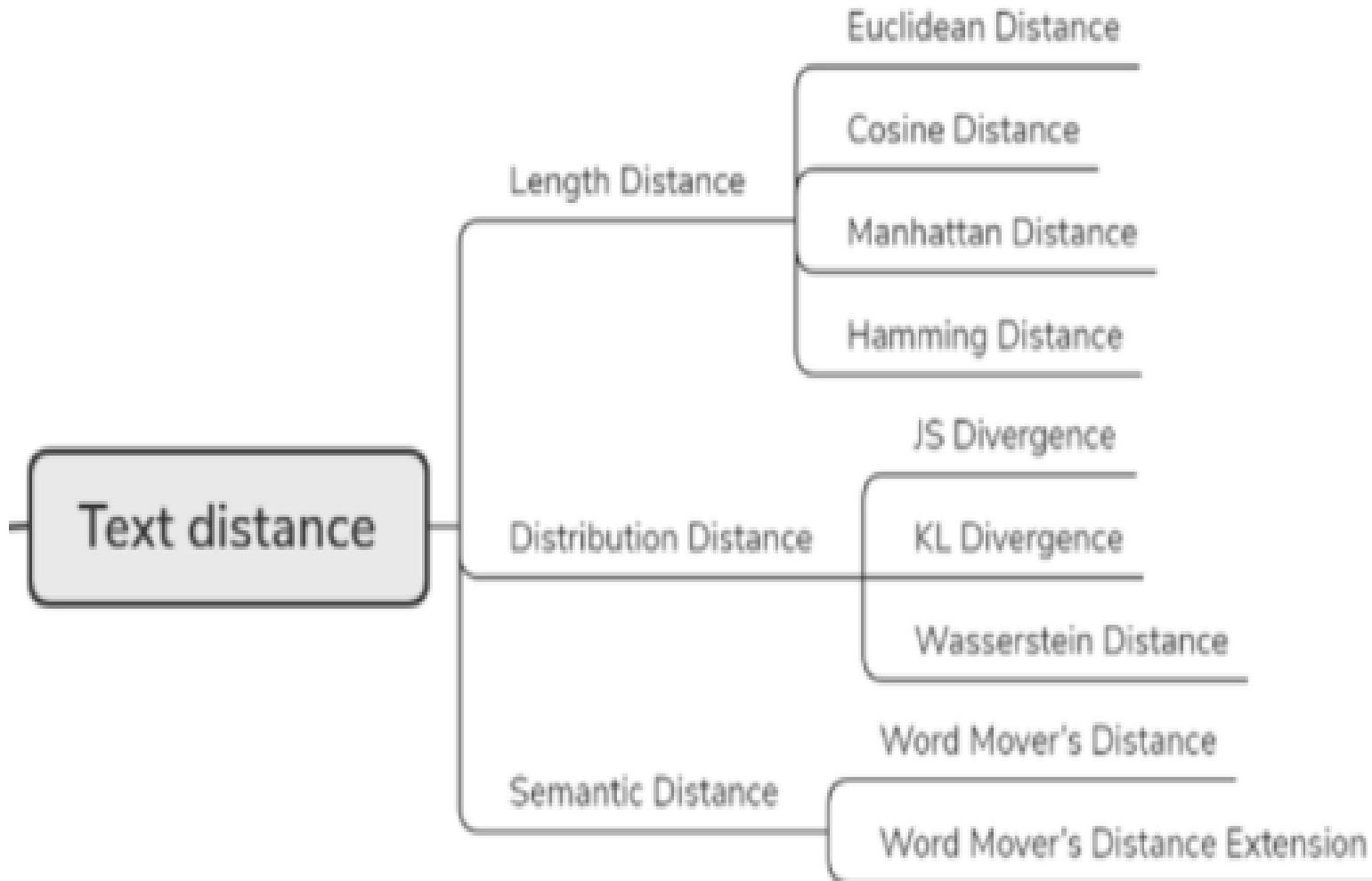
Text Similarity and computations

- Text similarity has to determine how ‘close’ two pieces of text are both in surface closeness [**lexical similarity**] and meaning [**semantic similarity**].
- The distance measures from one text to another text.

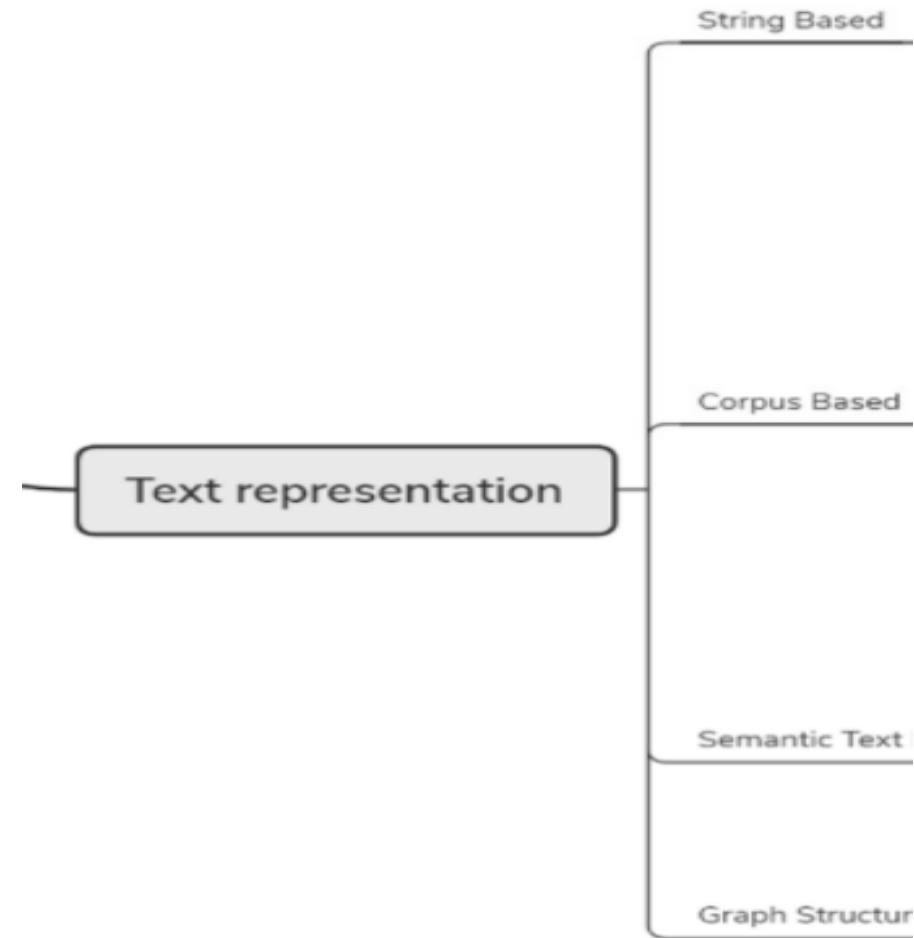
Similarity measures for textual data



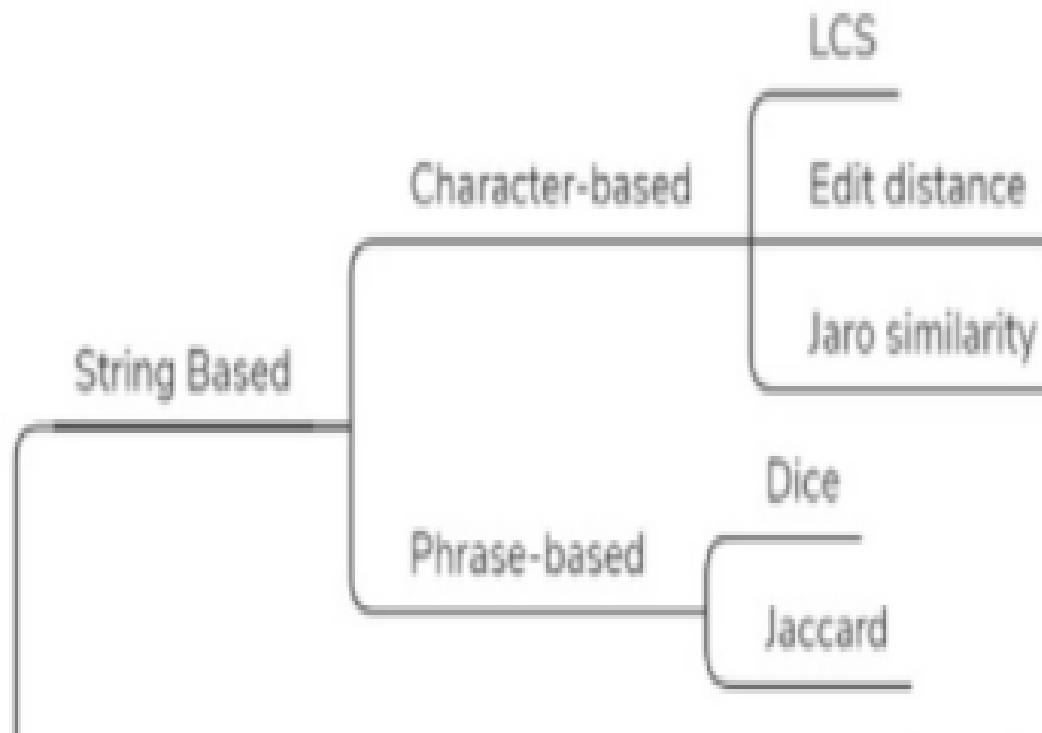
Text Distances



Text Representation



String Based



Text representation

- text matching, and graph-structure-baseText as numerical features that can be calculated directly.
- Texts can be similar in two ways lexically and semantically.
- The words that make up the text are similar lexically if they have a similar character sequence.
- Words are similar semantically if they have the same thing, are opposite of each other, used in the same way, used in the same context, and one is a type of another.
- Lexical similarity is introduced in this survey through different measurements of text representation, semantic similarity is introduced through the string-based method, corpus-based method, semantic d method.

Character-Based Text Representation

- A character-based similarity calculation is based on the similarity between characters in the text to express the similarity between texts.
- Three algorithms will be introduced:
 - LCS (longest common substring),
 - editing distance
 - Jaro similarity,

Jaro Distances

Jaro similarity

- The Jaro distance is a measure of edit distance between two strings; its inverse, called the *Jaro similarity*, is a measure of two strings' similarity: the higher the value, the more similar the strings are. The score is normalized such that **0** equates to no similarities and **1** is an exact match.

JARO Similarity Definition

- The Jaro similarity d_j of two given strings S_1 and S_2 is

$$d_j = \begin{cases} 0 & \text{if } m = 0 \\ \frac{1}{3} \left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m} \right) & \text{otherwise} \end{cases}$$

Where:

- m is the number of *matching characters*;
- t is half the number of *transpositions*.

JARO Similarity

- Two characters from S1 and S2 respectively, are considered *matching* only if they are the same and not farther apart than
$$\left\lfloor \frac{\max(|s_1|, |s_2|)}{2} \right\rfloor - 1 \text{ characters.}$$
- Each character of S1 is compared with all its matching characters in S2 .
- Each difference in position is half a *transposition*; that is, the number of transpositions is half the number of characters which are common to the two strings but occupy different positions in each one



Example

- Given 2 statements
 - Julie loves me more than Linda loves me
 - Jane likes me more than Julie loves me
- Find how similar these texts are, purely in terms of word counts (and ignoring word order)

SOLUTION

me	2	2
Jane	0	1
Julie	1	1
Linda	1	0
likes	0	1
loves	2	1
more	1	1
than	1	1

- The two vectors are, again:
- a: [2, 0, 1, 1, 0, 2, 1, 1]
- b: [2, 1, 1, 0, 1, 1, 1, 1]
- The cosine of the angle between them is about 0.822.
- These vectors are 8-dimensional. A virtue of using cosine similarity is clearly that it converts a question that is beyond human ability to visualise to one that can be. In this case you can think of this as the angle of about 35 degrees which is some 'distance' from zero or perfect agreement.

Jaro distance

- **Jaro Similarity** is the measure of similarity between two strings. The value of Jaro distance ranges from 0 to 1. where **1 means the strings are equal** and **0 means no similarity between two strings. It checks sequence similarity**
- Jaro distance measure is based on the **number or order of the characters between two strings which are common**;

$$\text{Jaro similarity} = \begin{cases} 0, & \text{if } m=0 \\ \frac{1}{3} \left(\frac{m}{|s1|} + \frac{m}{|s2|} + \frac{m-t}{m} \right), & \text{for } m \neq 0 \end{cases}$$

where:

- **m** is the **number of matching characters**
- **t** is **half the number of transpositions**
- **|s1|** and **|s2|** is the **length of string s1 and s2 respectively**
- **Condition for matching characters and getting number of transpositions:** Two characters from s1 and s2 are considered matching only if they are the same and not farther than $\left\lfloor \frac{\max(|s1|, |s2|)}{2} \right\rfloor - 1$ characters apart.

- For example, in comparing **CRATE** with **TRACE**, only 'R' 'A' 'E' are the **matching characters**, i.e. **m=3**. Although 'C', 'T' appear in both strings, they are farther apart than 1 (the result of $(5/2)\text{floor} - 1$). Therefore, **t=0** .
- Each character of **s_1** is compared with all its matching characters in **s_2**.The number of matching (but different sequence order) characters divided by 2 defines the number of transpositions.

EXAMPLE 1

- Let $s1="arnab"$, $s2="raanb"$, Calculate Jaro distance.
- Number of matching characters= 5 . Since Not farther than $(5/2)\text{floor}-1$.



- But the order is not the same, so the number of characters which are **not in order** is 4, so the **number of transpositions/2** is $4/2=2$.
- Therefore Jaro similarity can be calculated as follows:
 $\text{Jaro Similariy} = (1/3) * \{(5/5) + (5/5) + (5-2)/5\} = 0.86667$
- Jaro Distance =** $1-0.86667=0.13333$

EXAMPLE2

- Given S1=WINKLER and s2=WELFARE

W	I	N	K	L	E	R
W	E	L	F	A	R	E

- Matching characters=WLER and WLRE
- M=4, S1=7 S2=7
- Transposition=2/2=1
- Jaro Similarity= $(1/3) * \{(4/7) + (4/7) + (4-1)/4\} = 53/(3*28) = 53/84 = 0.63095$

W	L	E	R
W	L	R	E

EXAMPLE 2

1. Given the strings $s_1 = \text{"martha"}$ and $s_2 = \text{"marhta"}$. Calculate Jaro Similarity.
2. Given the strings $s_1 = \text{DWAYNE}$ and $s_2 = \text{DUANE}$. Calculate Jaro Similarity.
3. Given the strings $s_1 = \text{"CRATE"}$ and $s_2 = \text{"TRACE"}$. Calculate Jaro Similarity.

Solution

1. $m = 6$

$t = 2/2 = 1$ (2 couples of non matching characters, the 4-th and 5-th) { t/h ; h/t }

$|s_1| = 6$ and $|s_2| = 6$

- Jaro Similarity = $(\frac{1}{3}) (6/6 + 6/6 + (6-1)/6) = 17/18 = 0.944$

m	a	r	t	h	a
m	a	r	h	t	a

2. $m=4$

$t=0$. (In DwAyNE versus DuANE the matching letters are already in the same order D-A-N-E, so no transpositions are needed.)

$|s_1|=6$ and $|s_2|=5$

- Jaro Similarity = $(\frac{1}{3}) * (4/6 + 4/5 + (4-0)/4) = 37/45=0.822$

3. $m = 3$

$t = 0$ (only 'R' 'A' 'E' are the matching characters, i.e. $m=3$. Although 'C', 'T' appear in both strings, they are farther apart than 1 (the result of $(5/2)\text{floor} - 1$. Therefore, $t=0$.)

$|s_1| = 5$ and $|s_2| = 5$

- Jaro Similarity = $(\frac{1}{3}) * (3/5 + 3/5 + (3-0)/3) = 11/15=0.733333$

C	R	A	T	E
T	R	A	C	E

Jaro-Winkler Similarity

- Jaro – Winkler Similarity uses a prefix scale ‘p’ which **gives a more accurate answer when the strings have a common prefix up to a defined maximum length l.**
- Jaro Winkler similarity is defined as follows

$$Sw = Sj + P * L * (1 - Sj)$$

- where, **Sj**, is jaro similarity
- **Sw**, is jaro- winkler similarity
- **P** is the scaling factor (0.1 by default)
- **L** is the length of the matching prefix upto a maximum 4 characters
- The lower the Jaro–Winkler distance for two strings is, the more similar the strings are. The score is normalized such that 1 means an exact match and 0 means there is no similarity. The **Jaro–Winkler similarity** is the inversion
- Although often referred to as a *distance metric*, the Jaro–Winkler distance is not a metric in the mathematical sense of that term because it does not obey the triangle inequality.

EXAMPLE

- Let $s1="arnab"$, $s2="aranb"$. The Jaro similarity of two strings is 0.933333 (From the above calculation.)
- The length of matching prefix is 2(ar) and we take scaling factor as 0.1
- Substituting in the formula;
$$\text{Jaro-Winkler Similarity} = 0.9333333 + 0.1 * 2 * (1 - 0.9333333) = 0.946667$$

EXAMPLE

- “*martha*”/“*marhta*”
- prefix lenght of $I = 3$ (which refers to “*mar*”). We get to:
- $d_w = 0,944 + ((0,1^*3)(1-0,944)) = 0,944 + 0,3^*0,056 = 0,961$ Jaro-Winkler distance = 96,1%

Example

Given the strings s_1 DWAYNE and s_2 DUANE we find:

- $m = 4$
- $|s_1| = 6$
- $|s_2| = 5$
- $t = 0$

Solution : To find a Jaro score of given data

$$d_j = \frac{1}{3} \left(\frac{4}{6} + \frac{4}{5} + \frac{4 - 0}{4} \right) = 0.822$$

Lecture: 20 Similarity Measures for Textual Data

03/03/2022

Similarity Measures for Textual Data

1. Edit distance
2. n-Gram distance
3. Dissimilarity between attributes of mixed type



Agenda

- edit distance, n-Gram distance ,Dissimilarity between attributes of mixed type

EDIT DISTANCE APPLICATION

- Spell correction
 - The user typed “graffe”
Which is closest?
 - graf
 - graft
 - grail
 - giraffe
- Computational Biology
 - Align two sequences of nucleotides

AGGCTATCACCTGACCTCCAGGCCGATGCC
TAGCTATCACGACC CGCGGTGATTGCCCGAC

 - Resulting alignment:

- **AGGCTATCACCTGACCTCCAGGCCGA**--TGCCC---
TAG-CTATCAC--**GACC**GC--GGT**CGA**TTTGCCC**GAC**
- Also for Machine Translation, Information Extraction, Speech Recognition

EDIT DISTANCE APPLICATION

- Evaluating Machine Translation and speech recognition

R Spokesman confirms senior government adviser was shot

H Spokesman said the senior adviser was shot dead

S I

D

I

- Named Entity Extraction and Entity Coreference

- IBM Inc. announced today

- IBM profits

- Stanford President John Hennessy announced yesterday

- for Stanford University President John Hennessy

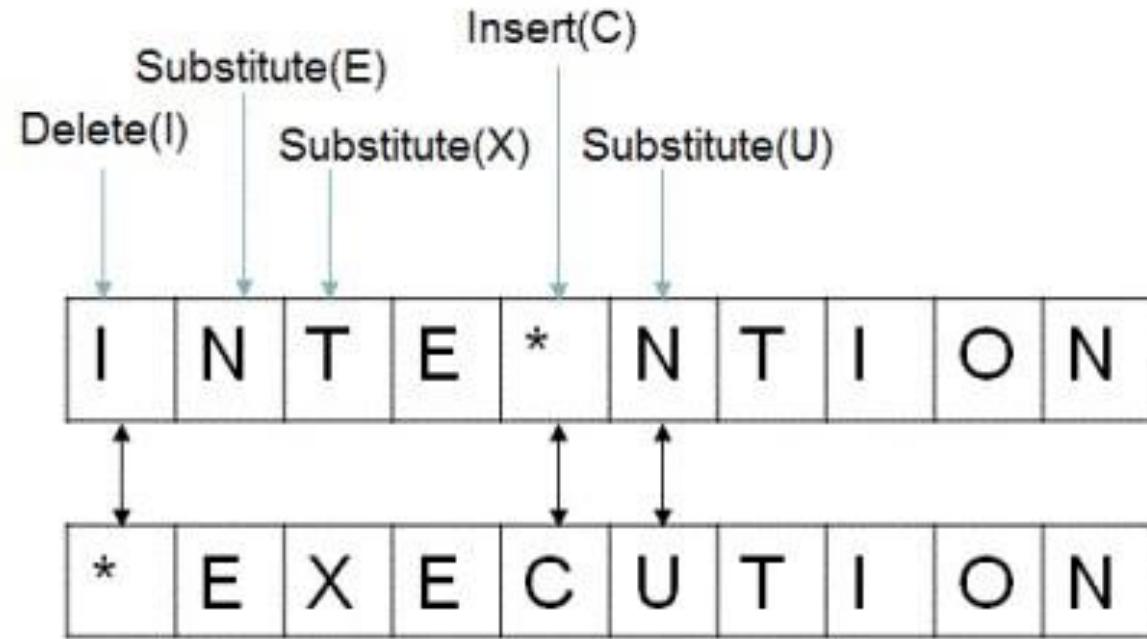


Edit distance

- **Minimum Edit distance** between two strings str1 and str2 is defined as the **minimum number of editing operations** like **insert, delete, substitute** required to transform **str1 into str2**.
- All the operations are of equal cost.
- **Edit distance** is a way of quantifying how dissimilar two strings are to one another by counting the minimum number of operations required to transform one string into the other.
- For example if str1 = "ab", str2 = "abc" then making an insert operation of character 'c' on str1 transforms str1 into str2. Therefore, edit distance between str1 and str2 is **1**.

EXAMPLE

- if str1 = "INTENTION" and str2 = "EXECUTION", then the minimum edit distance between str1 and str2 turns out to be 5 .
- If substitution cost is 2 then it becomes **Levenshtein Distance** which is 8.



EDIT DISTANCE PROBLEM AS AN OPTIMAL SUBSTRUCTURE using Dynamic Programmin g

- The Edit distance problem has an optimal substructure(problem can be broken down into smaller, simple , which can be broken down into yet simpler subproblems, and so on, until, finally, the solution becomes trivial).

$$\text{Dist}[i][j] = \begin{cases} \text{when } X[i-1] == Y[j-1] \\ \quad \text{dist}[i - 1][j - 1] \\ \text{when } X[i-1] != Y[j-1] \\ \quad 1 + \text{Min} \left\{ \begin{array}{l} \text{dist}[i-1][j], \\ \text{dist}[i][j - 1], \\ \text{dist}[i - 1][j - 1] \end{array} \right\} \end{cases}$$

Why Backtrackin g is needed?

- Edit distance isn't sufficient
 - We often need to align each character of the two strings to each other
- We do this by keeping a “backtrace”
- Every time we enter a cell, remember where we came from
- When we reach the end,
 - Trace back the Path from The Upper Right Corner To Read Off The alignment

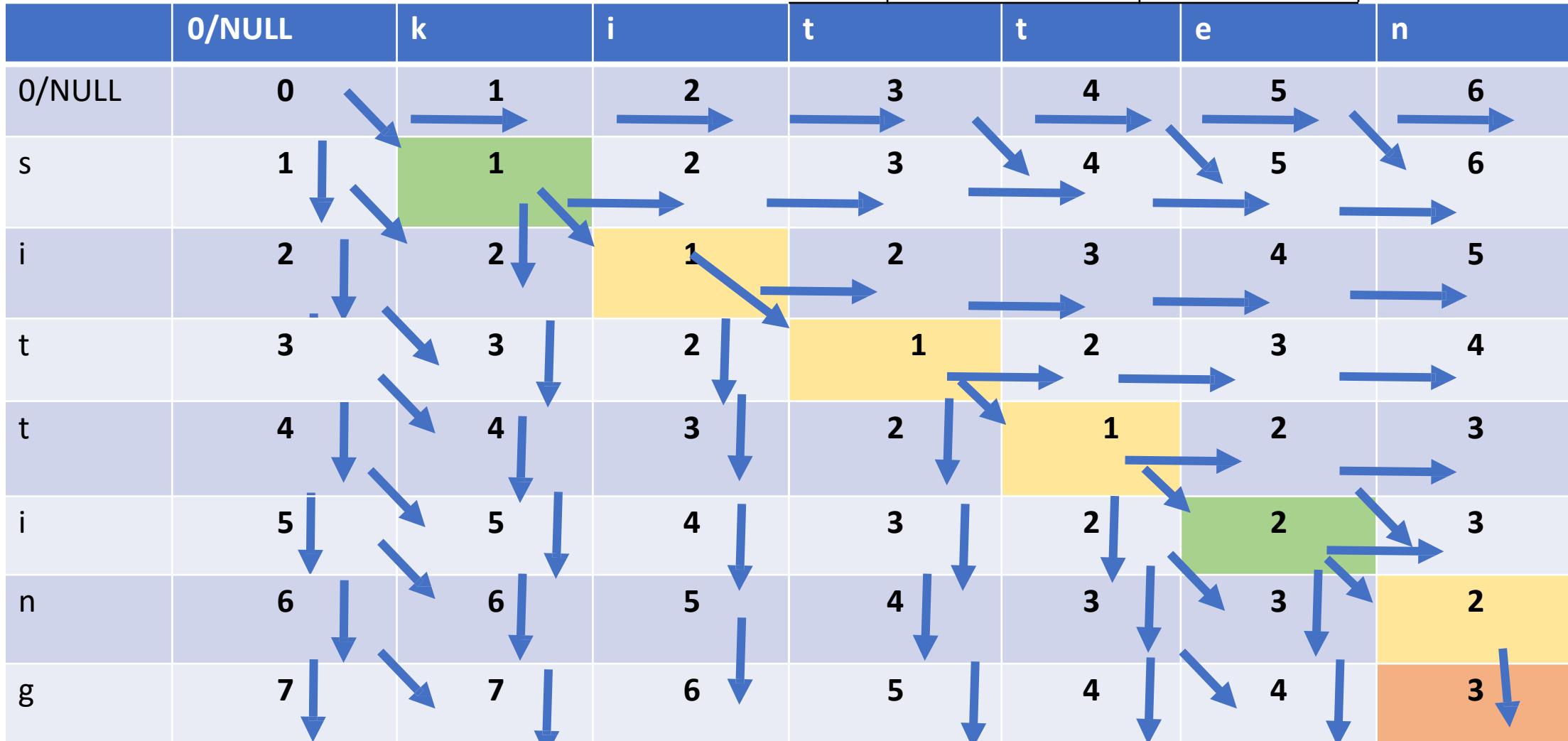
If($r==c$) just copy the diagonal elements

substitute	delete
insert	$\text{Min}(\text{substitute}, \text{del}, \text{ete}, \text{insert}) + 1$

	0/NULL	k	i	t	t	e	n
0/NULL	0	1	2	3	4	5	6
s	1						
i	2						
t	3						
t	4						
e	5						
n	6						
g	7						

If($r==c$) just copy the diagonal elements

substitute		delete	
insert		Min(substitute, delete, insert) + 1	



CHALLENGE

- Compute Edit distance table for **STR1**: Tamming test and **STR 2**: Taming text

		t	a	m	m	i	n	g		t	e	s	t
	0	1	2	3	4	5	6	7	8	9	10	11	12
t	1	0	1	2	3	4	5	6	7	8	9	10	11
a	2	1	0	1	2	3	4	5	6	7	8	9	10
m	3	2	1	0	1	2	3	4	5	6	7	8	9
i	4	3	2	1	1	1	2	3	4	5	6	7	8
n	5	4	3	2	2	2	1	2	3	4	5	6	7
g	6	5	4	3	3	3	2	1	2	3	4	5	6
	7	6	5	4	4	4	3	2	1	2	3	4	5
t	8	7	6	5	5	5	4	3	2	1	2	3	4
e	9	8	7	6	6	6	5	4	3	2	1	2	3
x	10	8	8	7	7	7	6	5	4	3	2	2	3
t	11	8	9	8	8	8	7	6	5	3	3	3	2

Dynamic Programming for Levenshtein Distance

Initialization

$$D(i, 0) = i$$

$$D(0, j) = j$$

$$D[i, j] = \min \begin{cases} D[i - 1, j] + \text{del-cost}(\text{source}[i]) \\ D[i, j - 1] + \text{ins-cost}(\text{target}[j]) \\ D[i - 1, j - 1] + \text{sub-cost}(\text{source}[i], \text{target}[j]) \end{cases}$$

Recurrence Relation:

For each $i = 1 \dots M$

For each $j = 1 \dots N$

$$D(i, j) = \min \begin{cases} D(i - 1, j) + 1 \\ D(i, j - 1) + 1 \\ D(i - 1, j - 1) + 2; & \text{if } X(i) \neq Y(j) \\ 0; & \text{if } X(i) = Y(j) \end{cases}$$

Termination:

$D(N, M)$ is distance

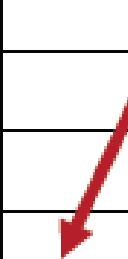
EXAMPLE

N	9										
O	8										
I	7										
T	6										
N	5										
E	4										
T	3										
N	2										
I	1										
#	0	1	2	3	4	5	6	7	8	9	
	#	E	X	E	C	U	T	I	O	N	

EXAMPLE

N	9									
O	8									
I	7									
T	6									
N	5									
E	4									
T	3									
N	2									
I	1									
#	0	1	2	3	4	5	6	7	8	9
	#	E	X	E	C	U	T	I	O	N

$D(i,j) = \min \begin{cases} D(i-1,j) + 1 \\ D(i,j-1) + 1 \\ D(i-1,j-1) + \begin{cases} 2; & \text{if } S_1(i) \neq S_2(j) \\ 0; & \text{if } S_1(i) = S_2(j) \end{cases} \end{cases}$



EXAMPLE

N	9	8	9	10	11	12	11	10	9	8
O	8	7	8	9	10	11	10	9	8	9
I	7	6	7	8	9	10	9	8	9	10
T	6	5	6	7	8	9	8	9	10	11
N	5	4	5	6	7	8	9	10	11	10
E	4	3	4	5	6	7	8	9	10	9
T	3	4	5	6	7	8	7	8	9	8
N	2	3	4	5	6	7	8	7	8	7
I	1	2	3	4	5	6	7	6	7	8
#	0	1	2	3	4	5	6	7	8	9
	#	E	X	E	C	U	T	I	O	N

Adding Backtrack

Base conditions:

$$D(i, 0) = i$$

$$D(0, j) = j$$

Termination:

D(N,M) is distance

Recurrence Relation:

For each $i = 1 \dots M$

For each $j = 1 \dots N$

$$D(i, j) = \min \begin{cases} D(i-1, j) + 1 & \text{deletion} \\ D(i, j-1) + 1 & \text{insertion} \\ D(i-1, j-1) + 2; & \begin{cases} \text{if } X(i) \neq Y(j) & \text{substitution} \\ 0; & \text{if } X(i) = Y(j) \end{cases} \\ \text{LEFT} & \text{insertion} \\ \text{DOWN} & \text{deletion} \\ \text{DIAG} & \text{substitution} \end{cases}$$

BACKTRACKING LEVENSHTEIN

n	9	↓ 8	↙ ↘ 9	↙ ↘ 10	↙ ↘ 11	↙ ↘ 12	↓ 11	↓ 10	↓ 9	↙ 8	
o	8	↓ 7	↙ ↘ 8	↙ ↘ 9	↙ ↘ 10	↙ ↘ 11	↓ 10	↓ 9	↙ 8	← 9	
i	7	↓ 6	↙ ↘ 7	↙ ↘ 8	↙ ↘ 9	↙ ↘ 10	↓ 9	↙ 8	← 9	← 10	
t	6	↓ 5	↙ ↘ 6	↙ ↘ 7	↙ ↘ 8	↙ ↘ 9	↙ 8	← 9	← 10	← 11	
n	5	↓ 4	↙ ↘ 5	↙ ↘ 6	↙ ↘ 7	↙ ↘ 8	↙ ↘ 9	↙ ↘ 10	↙ ↘ 11	↙ ↘ 10	
e	4	↙ 3	← 4	↙ ↘ 5	← 6	← 7	← 8	↙ ↘ 9	↙ ↘ 10	↓ 9	
t	3	↙ ↘ 4	↙ ↘ 5	↙ ↘ 6	↙ ↘ 7	↙ ↘ 8	↙ 7	← 8	↙ ↘ 9	↓ 8	
n	2	↙ ↘ 3	↙ ↘ 4	↙ ↘ 5	↙ ↘ 6	↙ ↘ 7	↙ ↘ 8	↓ 7	↙ ↘ 8	↙ 7	
i	1	↙ ↘ 2	↙ ↘ 3	↙ ↘ 4	↙ ↘ 5	↙ ↘ 6	↙ ↘ 7	↙ 6	← 7	← 8	
#	0	1	2	3	4	5	6	7	8	9	
	#	e	x	e	c	u	t	i	o	n	

Types of Edit Distance

- Different types of edit distance allow different sets of string operations. For instance:
 - **Levenshtein distance** allows **deletion, insertion and substitution**.
 - **Longest common subsequence (LCS)** distance allows only **insertion and deletion**, not substitution.
 - **Hamming distance** allows only **substitution**, hence, it only applies to strings of the same length.
 - **Jaro distance** allows only **transposition**.

N-gram Distances

n-Gram edit distance

- An **n-gram** is a **contiguous sequence of n items** from a given sample of text or speech. The items can be phonemes, syllables, letters, words or base pairs according to the application.
- An n-gram can be obtained by splitting a string in sequences with the length n, ie substrings of length N are named "N-grams".
- **Example:** string “abcde”
 - **bigrams** are: ab, bc, cd, and de
 - **trigrams** will be: abc, bcd, and cde
 - **4-grams** will be abcd, and bcde.
- **There are different variants of N-gram distance:**
 - **Dice N-gram Matching Method**
 - **Generalized N-gram Matching for string matching**
 - **Bigram method**
 - **Trigram Method**

Dice N-gram Matching

- **Dice similarity** is the most popular N-Gram method.
- The measures are defined as the **ratio of the number of n-grams that are shared by two strings and the total number of n-grams in both strings**.
- The coefficient value varies between zero and one. If two terms have no characters in common then the coefficient value is zero. On the other hand, if they are identical, the coefficient value will be one.
- For two string X and Y, the Dice coefficient is measured as

$$d(X, Y) = \frac{2(n - gram(X \cap Y))}{(n - gram(X)) + (n - gram(Y))}$$

- where $n\text{-grams}(X)$ is a multi-set of letter n-grams in X .
- Dice coefficient with bigrams (DICE) is a particularly popular word similarity measure

Dice N-gram Matching

- DICE(Zantac, Contac)
- Zantac: za, an, nt, ta, ac
- Contac: co, on, nt, ta, ac
- $d(X, Y) = \frac{2(n - gram(X \cap Y))}{(n - gram(X)) + (n - gram(Y))}$ = $(2 \cdot 3)/(5 + 5) = 6/10 = 0.6$

• Problems of Dice N-gram

- **Low resolution:** it often fails to detect any similarity between strings that are very much alike.
Example: Verelan/Virilon have no n-grams in common.
- It can return the **maximum similarity value of 1** for **strings that are non-identical**. Example, both Xanex and Nexan are composed of the same set of bigrams: {an,ex,ne,xa}.
- It often associates n-grams that occur in radically different word positions, as in the pair Voltaren/Tramadol.

N-gram Distance

- N-gram distance:

$$\text{Sim}(s_1, s_2) = \frac{1}{N-n+1} \sum_{i=0}^{N-n+1} (h(i))$$

- Where $N = \max(N(s_1), N(s_2))$
- $n=2$ for bigram, 3 for trigram, 4 for 4-grams
- $h(i) = 1$ if n-element subsequence beginning from position i in s_1 appears in s_2 $h(i) = 0$ otherwise

Bigram and Trigram

- Bigram

$$\text{sim}(s_1, s_2) = \frac{1}{N-2+1} \sum_{i=0}^{N-2+1} h(i) = \frac{1}{N-1} \sum_{i=0}^{N-1} h(i)$$

- Trigram

$$\text{sim}(s_1, s_2) = \frac{1}{N-3+1} \sum_{i=0}^{N-3+1} h(i) = \frac{1}{N-2} \sum_{i=0}^{N-2} h(i)$$

Generalized N-gram Matching

- Generalized n-gram matching was introduced by Niewiadomski

$$sim(s_1, s_2) = f(n_1, n_2) \sum_{i=n_1}^{n_2} \sum_{j=1}^{N-n+1} h(i, j)$$

- where $f(n_1, n_2) = \frac{2}{(N-n_1+1)(N-n_2+2)-(N-n_2+1)(N-n_1)} = 2/(N^2+N)$
- denotes the number of possible substrings not shorter than n_1 and not longer than n_2 in s_1

EXAMPLE

1. Let $s_1 = \text{ELOQUENTLY}$, $s_2 = \text{INELOQUENT}$. $N(s_1) = 10$ and $N(s_2) = 10$. Calculate the 4 n-gram distances.

SOLUTION:

s_2 occurs in the substring of s_1 as follows:

1-element E, L, O, Q, U, E, N, T = 8

2 element EL, LO, OQ, QU, UE, EN, NT, = 7

3 element ELO, LOQ, OQU, QUE, UEN, ENT = 6

4 element ELOQ, LOQU, OQUE, QUEN, UENT = 5

5 -element ELOQU, LOQUE, OQUEN, QUENT = 4

6 element ELOQUE, LOQUEN, OQUENT = 3

7 element ELOQUEN, LOQUENT = 2

8 -element ELOQUENT = 1

EXAMPLE

1. Generalized n-gram matching

$$\text{sim}(s_1, s_2) = \frac{2}{N^2+N} \sum_{i=1}^N \sum_{j=1}^{N-i+1} h(i, j) = \frac{2}{10^2+10} \times \frac{8+7+6+5+4+3+2+1}{1} = \frac{2*36}{110} = 0.65$$

$$\text{Bigram} = \text{sim}(s_1, s_2) = \frac{1}{N-n+1} \sum_{i=0}^{N-n+1} h(i) = \frac{1}{10-1} \times \frac{7}{1} = \frac{7}{9} = 0.77$$

$$\text{Trigram} = \text{sim}(s_1, s_2) = \frac{1}{N-n+1} \sum_{i=0}^{N-n+1} h(i) = \frac{1}{10-2} \times 6 = \frac{6}{8} = 0.75$$

$$\text{Dice's Coefficient} = d(X, Y) = \frac{2(n\text{-gram}(X \cap Y))}{n\text{-gram}(X) + n\text{-gram}(Y)} = \frac{2(7)}{9+9} = \frac{14}{18} = 0.77$$

EXAMPLE 2

- Let $s_1 = \text{PROGRAMMER}$, $s_2 = \text{PROGRAMMING}$. $N(s_1) = 10$ and $N(s_2) = 11$, $\max\{N(s_1), N(s_2)\} = 11$. Calculate the 4 n-gram distances.

1. Generalized n-gram matching

$$sim(s_1, s_2) = \frac{2}{N^2+N} \sum_{i=1}^N \sum_{j=1}^{N-i+1} h(i, j) = \frac{2}{11^2+11} \times \frac{9+7+6+5+4+3+2+1}{1} = 2*36/132$$

$$72/132=0.545$$

$$2. \text{ Dice's Coefficient } d(X, Y) = \frac{2(n\text{-gram}(X \cap Y))}{n\text{-gram}(X) + (n\text{-gram}(Y))} = \frac{2(7)}{9+10} = \frac{14}{19} = 0.74$$

$$3. \text{ Bigram } sim(s_1, s_2) = \frac{1}{N-n+1} \sum_{i=0}^{N-n+1} h(i) = \frac{1}{11-1} \times \frac{7}{1} = \frac{7}{10} = 0.70$$

$$4. \text{ Trigram } sim(s_1, s_2) = \frac{1}{N-n+1} \sum_{i=0}^{N-n+1} h(i) = \frac{1}{11-2} \times \frac{6}{1} = \frac{6}{9} = 0.64$$

SOLUTION

- s_2 occurs in the substring of s_1 as follows:

1 element P, R, O, G, R, A, M, M = 8

2 element PR, RO, OG, GR, RA, AM, MM = 7

3 element PRO, ROG, OGR, GRA, RAM, AMM = 6

4 element PROG, ROGR, OGRA, GRAM, RAMM = 5

5 -element PROGR, ROGRA, OGRAM, GRAMM = 4

6 element PROGRA, ROGRAM, OGRAMM = 3

7 element PROOGRAM, ROGRAMM = 2

8 -element PROGRAMM = 1



Dissimilarity Between Attributes of mixed type

Dissimilarity Between Attributes of mixed type

- A database may contain different types of variables **interval-scaled, symmetric binary, asymmetric binary, nominal, and ordinal**

Approach 1

- compute the similarity between each attribute separately and then combine these attribute using a method that results in a similarity between 0 and 1. Combine the different variables into a single dissimilarity matrix, bringing all of the meaningful variables onto a common scale of the interval [0.0, 1.0].
- Suppose that the data set contains p variables of mixed type. The dissimilarity $d(i, j)$ between objects i and j is defined as

$$d(i, j) = \frac{\sum_{f=1}^p \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^p \delta_{ij}^{(f)}}$$

- $\delta_{ij}^{(f)} = 0$
 - if either (1) x_{if} or x_{jf} is **missing** (i.e., there is no measurement of variable f for object i or object j),
 - or (2) $x_{if} = x_{jf} = 0$ and variable f is **asymmetric binary**;
- otherwise $\delta_{ij}^{(f)} = 1$

Dissimilarity Between Attributes of mixed type

- The contribution of variable f to the dissimilarity between i and j, that is, $d_{ij}(f)$ is computed dependent on its type:
 - If f is **interval-based/numeric**:
 - use the normalized distance so that the values map to the interval [0.0,1.0].

$$d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{\max_h x_{hf} - \min_h x_{hf}}$$

- If f is **binary or categorical**:
 - $d_{ij}(f) = 0$ if $x_{if} = x_{jf}$, or $d_{ij}(f) = 1$ otherwise
- If f is **ordinal**:
 - compute ranks r_{if} and $z_{if} = r_{if} - 1/M_f - 1$, and treat z_{if} as numeric

EXAMPLE

Table 2.2: A sample data table containing attributes of mixed type.

<i>object identifier</i>	<i>test-1</i> (nominal)	<i>test-2</i> (ordinal)	<i>test-3</i> (numeric)
1	code-A	excellent	45
2	code-B	fair	22
3	code-C	good	64
4	code-A	excellent	28

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

Dissimilarity matrix for test-1

$$\begin{bmatrix} 0 & & & \\ 1.0 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1.0 & 0.5 & 0 \end{bmatrix}$$

Dissimilarity matrix for test-2

- use the dissimilarity matrices obtained for test -1 and test -2
- Compute the dissimilarity matrix for the third attribute, test-3 (which is numeric)

SOLUTION

- compute the dissimilarity matrix for the third attribute, test-3 (which is numeric)
- $\max_{x \in h} x = 64$ and $\min_{x \in h} x = 22$. The difference between the two is used to normalize the values of the dissimilarity matrix(eg: $d_{21} = (45 - 22) / (64 - 22) = 0.55$). Find all the d_{ij} values for attribute test-3
- The resulting dissimilarity matrix for test -3 is:

$$\begin{bmatrix} 0 \\ 0.55 & 0 \\ 0.45 & 1.00 & 0 \\ 0.40 & 0.14 & 0.86 & 0 \end{bmatrix}$$

SOLUTION

- Now compute the dissimilarity matrices for the three attributes: for each of the three attributes, the indicator $d_{au}(f)$ $ij = 1$
- Now compute each $d(i,j)$: e.g.

$$d(i,j) = \frac{\sum_{f=1}^p \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^p \delta_{ij}^{(f)}}$$

$$d(3,1) = \frac{1(1)+1(0.50)+1(0.45)}{3} = 0.65,$$

- The resulting dissimilarity matrix is

- As a result:
 - objects 1 and 4 are the most similar
 - objects 1 and 2 are the least similar

$$\begin{bmatrix} 0 & & & \\ 0.85 & 0 & & \\ 0.65 & 0.83 & 0 & \\ 0.13 & 0.71 & 0.79 & 0 \end{bmatrix}$$

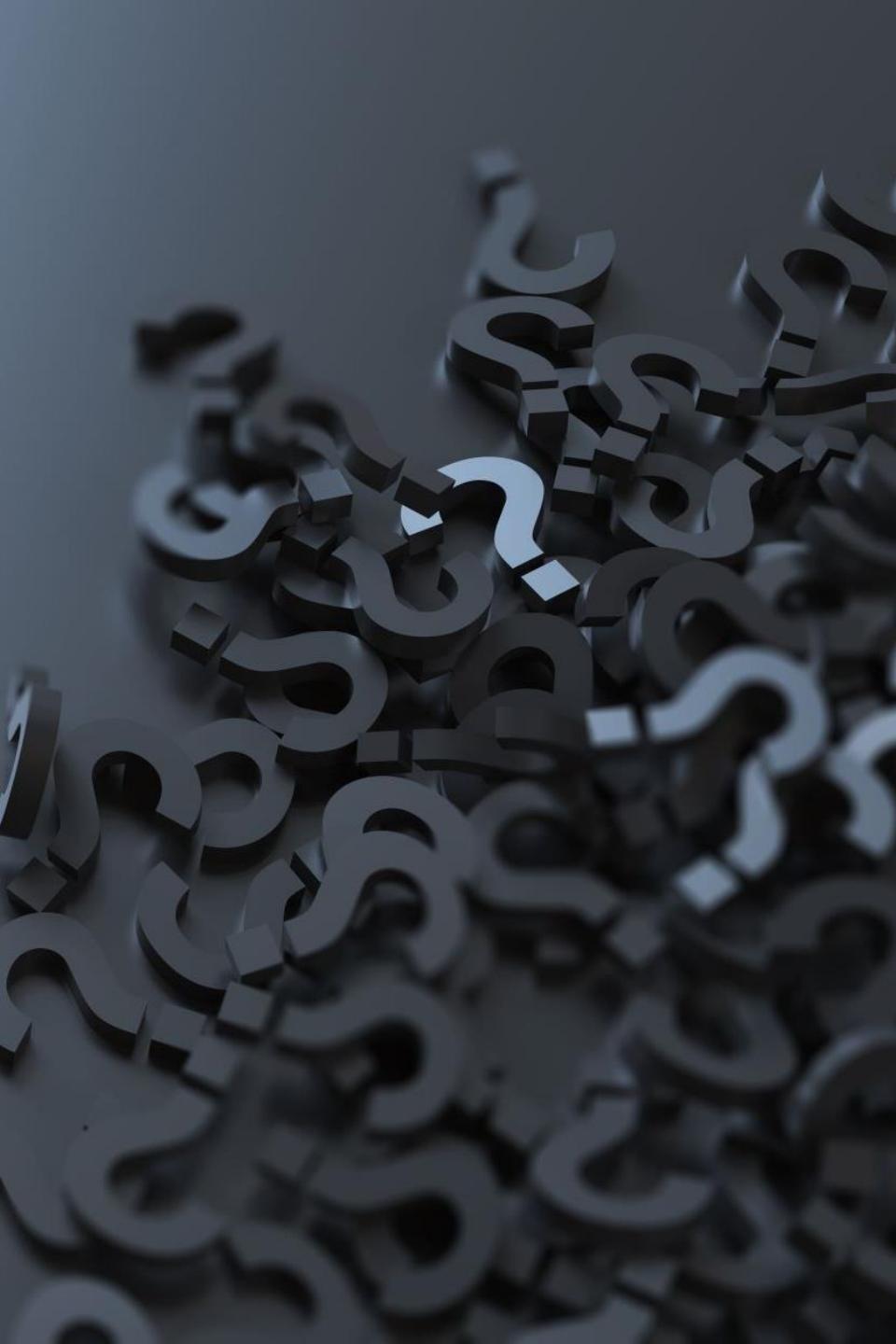
Proximity Measures for Interval-scaled Attributes

- Same as that of Numeric attributes

Proximity Measures for ratio-Scaled Attributes

Ratio-scaled attributes

- Numeric attributes, but unlike interval-scaled attributes, their scales are exponential,
- There are three methods :
 - treat them like interval-scaled variables — *not a good choice!(scale may be distorted)*
 - apply logarithmic transformation ($y_{if} = \log(x_{if})$) and then treat it as an interval-scaled attribute
 - treat them as continuous ordinal data treat their ranks as interval-scaled.



Similarity features of Text

- **Matching** – How many words or characters match between two inputs. Ex: “abc” and “abz” have 2 characters matching but 1 different.
- **Sequence** – Are matching characters in same order or sequence? Ex: “abc” and “abz” have matching characters ‘ab’ in same sequence. “abc” and “zba” have matching ‘ab’ characters but not in same sequence.
- **Sound** – Some spellings might be different, but when we pronounce them they might sound same. That might be a criteria to categorize them as similar.



Thank You