



DATA ANALYTICS

UE21CS342AA2

UNIT-1

**Lecture 1 : Introduction to DA , Data
Sources and Representations**

Gowri Srinivasa

Department of Computer Science and Engineering

Data Analytics

Unit 1

Lecture 1 : Introduction to Data Analytics , Data Sources and Representations

Slides excerpted from: U. Dinesh Kumar,
“Business Analytics”, Wiley, 2nd Edition 2022

Gowri Srinivasa

Department of Computer Science and Engineering

Slides collated by:

Nishanth M S, CSE 2023, PES University
nishanthmsathish.23@gmail.com

Harshitha Srikanth, VII CSE, PES University
harshithasrikanth13@gmail.com

With grateful thanks for contribution of slides to:
Dr. Mamatha H R, Professor at the Department of CSE, PESU

Course Instructors



Dr. Gowri Srinivasa



Dr. Bharathi R.



Dr. K. S. Nagegowda



Prof. Suresh Jamadagni



Prof. Srinivas K.S.



Dr. Sujatha R. Upadhyaya



Prof. Bhaskarjyoti Das



Dr. Jyothi R.



Dr. Prajwala T.R.



Prof. Lakshmeesha



Dr. Sudeepa Roy Dey

Evaluation Policy (tentative)

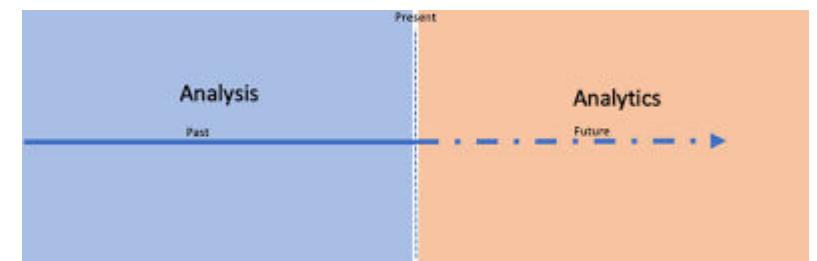
Component	Description	Weight
ISA 1 CBT	40 marks scaled to 20 (Unit 1 and Unit 2)	20
ISA-2 CBT	40 marks scaled to 20 (Unit 3 and Unit 4)	20
Experiential Learning: Worksheets (Lab + Assignment) + Hackathon	<ul style="list-style-type: none"> Worksheets/ assignments (1 per unit, with multiple parts) Hackathon (1 for the entire course): 5 marks 2 mark for participation + 2 marks for working + 2 mark for finishing in the top 30% of the leaderboard: 2 marks (for those not in the top 30%, class participation (attendance + questions asked, answered) in the lecture hours + invited talk can count towards the 1 mark)	4 6
ESA – pen and paper	100 marks scaled to 50	50

What is data analytics?

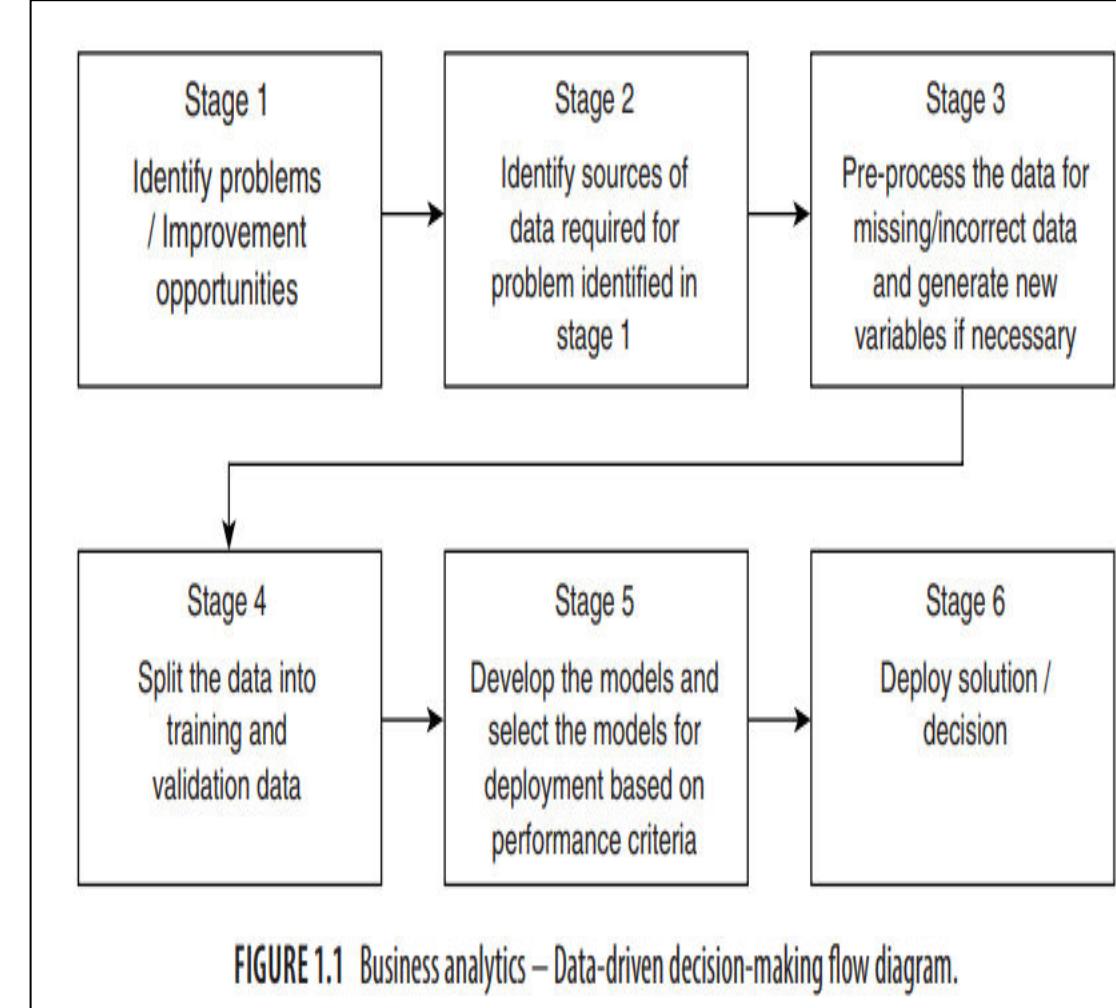
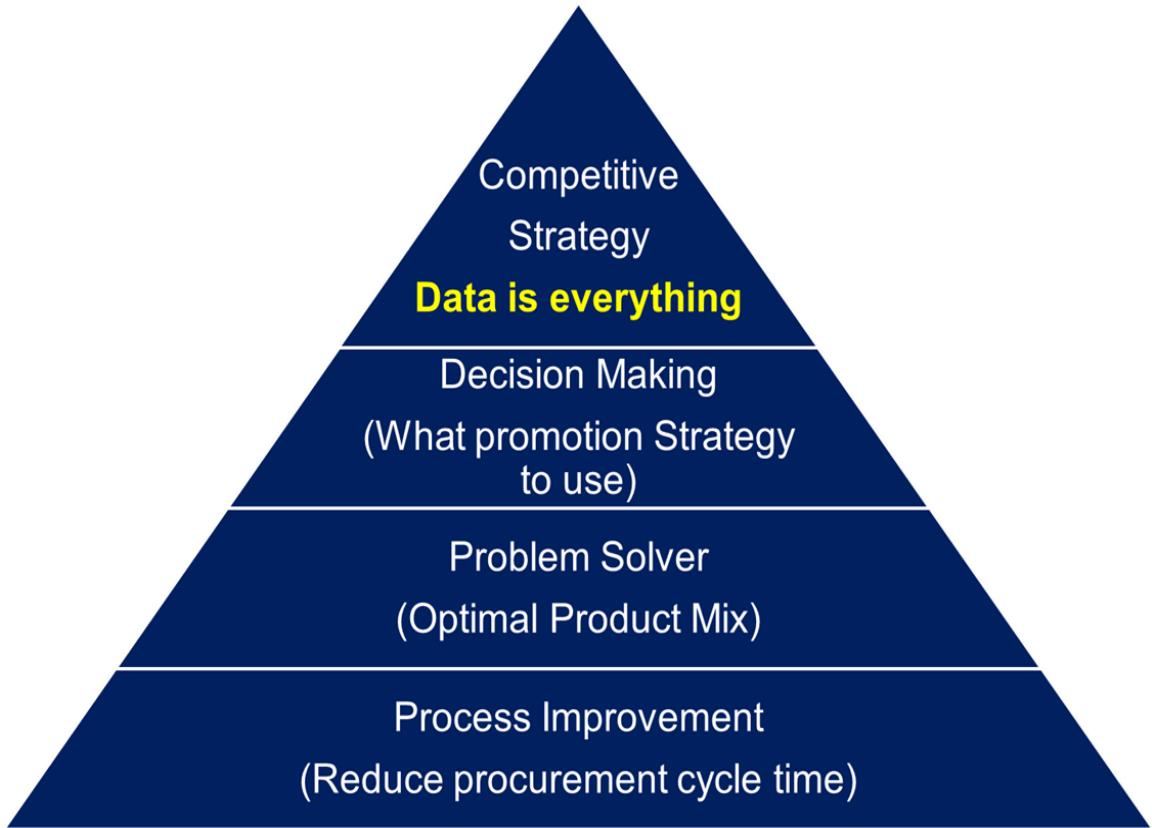
- The science of examining raw data to **elicit patterns, develop insights , and draw conclusions** to help take a **business decision**.

- The need :
Business decisions are very complex. There exist several alternate solutions, complex interdependent factors and lack of available time to take a decision.

- Analysis vs analytics
 - Analysis – Examining and understanding past data.
 - Analytics – Analysis + forecasting (or predictive modeling).



Pyramid of analytics applications and Data driven decision making



4 types of Data Analytics

Value

Prescriptive

Predictive

Diagnostic

Descriptive

What is the data telling you?

Descriptive: What's happening in my business?

- Comprehensive, accurate and live data
- Effective visualisation

Diagnostic: Why is it happening?

- Ability to drill down to the root-cause
- Ability to isolate all confounding information

Predictive: What's likely to happen?

- Business strategies have remained fairly consistent over time
- Historical patterns being used to predict specific outcomes using algorithms
- Decisions are automated using algorithms and technology

Prescriptive: What do I need to do?

- Recommended actions and strategies based on champion / challenger testing strategy outcomes
- Applying advanced analytical techniques to make specific recommendations

Complexity

Innovative ways to summarize data

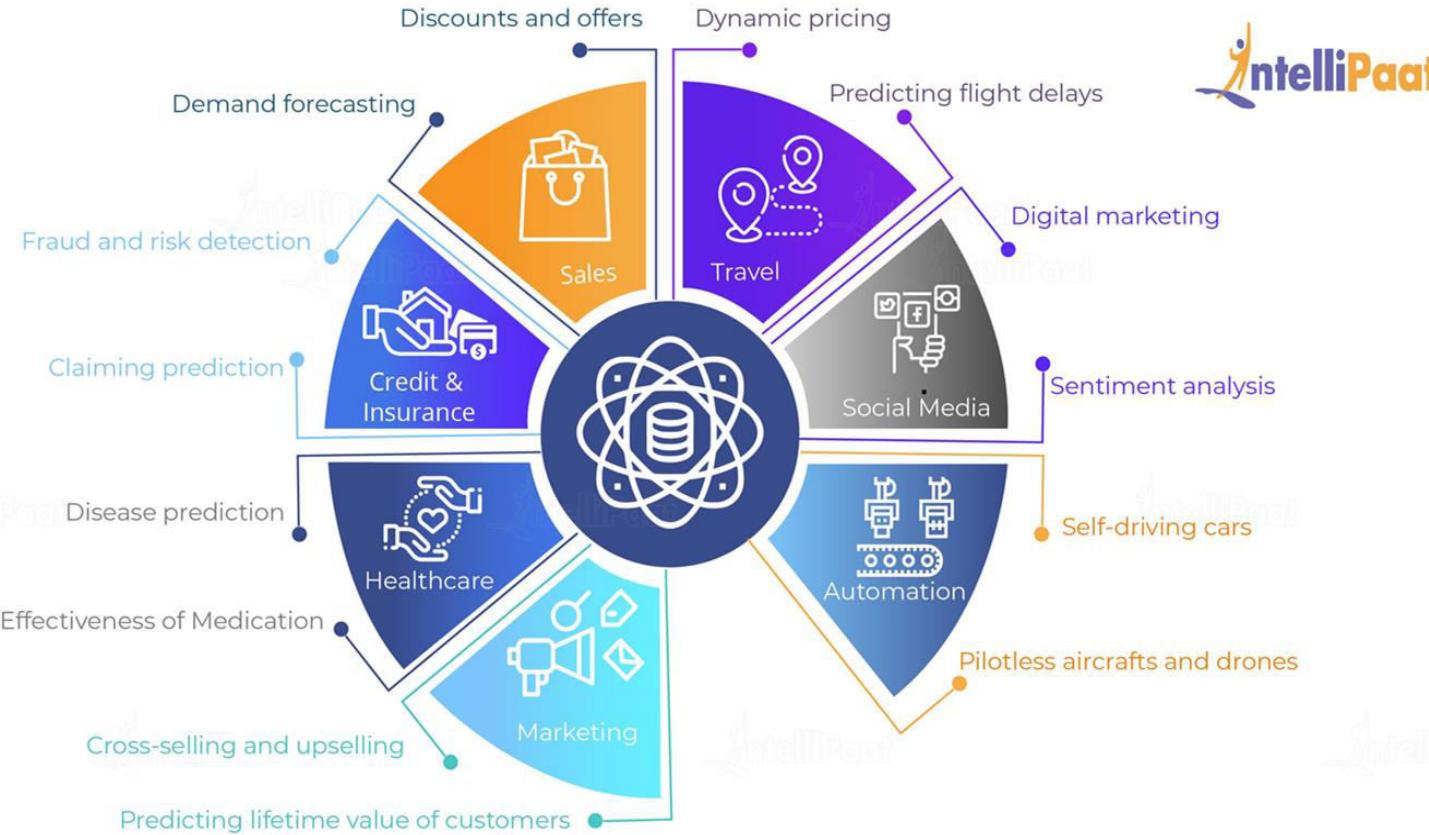
- Understanding trends in past data
- Hidden facts/trends

The primary aim of diagnostic analytics is to determine the causes of observed outcomes. For example, if a company's sales dropped in the previous quarter (an insight obtained from descriptive analytics), diagnostic analytics would seek to find out why that decline occurred.

Eg: Netflix , amazon predictive analytics to recommend movies ,product.

Why is it important?

Data Analytics is used in all these application areas...

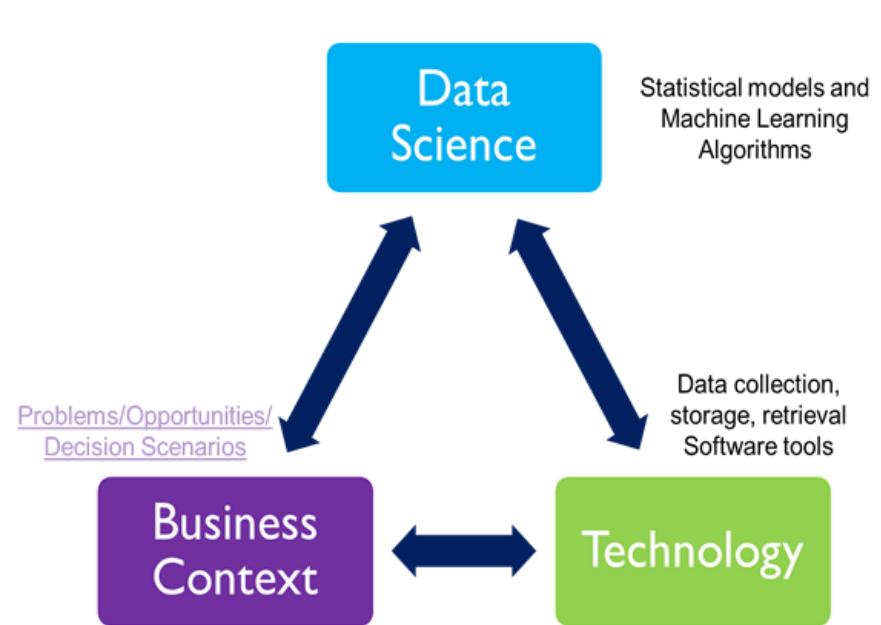


... and more!

DATA ANALYTICS

Few examples

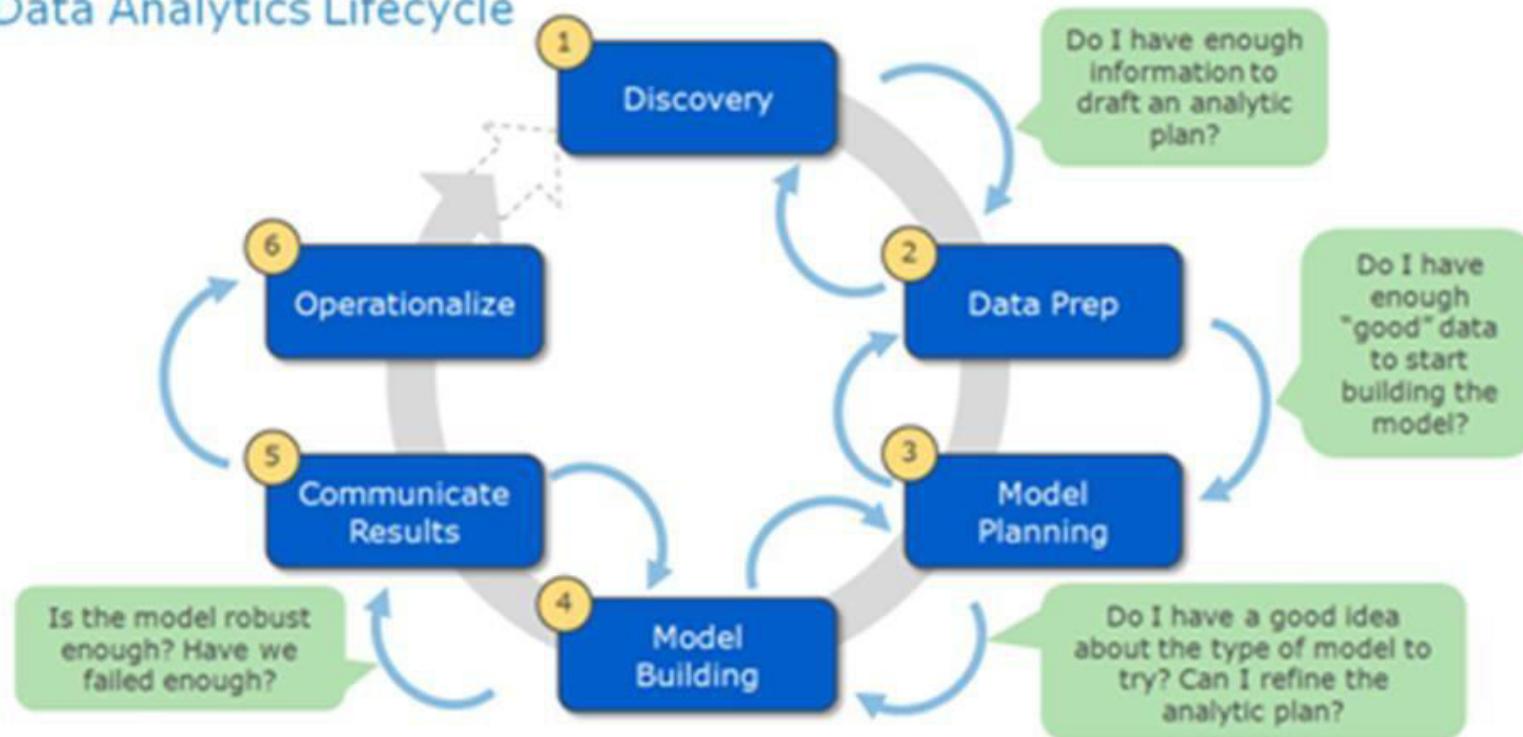
- Banking : To reduce cheque clearance time, in determining loan approval and interest rate.
- E-Commerce : To analyze buyer behavior to plan inventory and recommend products.
- Retail stores : Shelf space allocation to drive the profits up.
- OTT Platforms : Recommend content a user would like.



DATA ANALYTICS

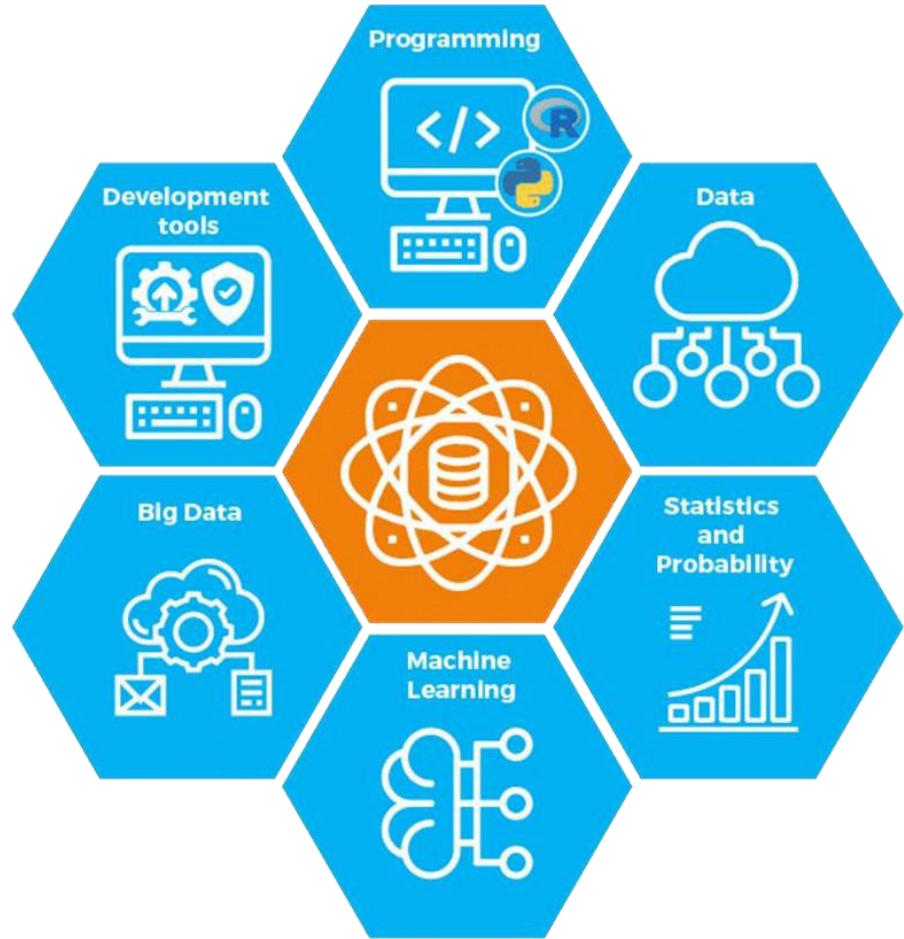
Lifecycle

Data Analytics Lifecycle



DATA ANALYTICS

Skills Required



And...
a secret ingredient



Intuition or
deductive reasoning
and domain knowledge

Case Study

Indian online grocery store bigbasket.com

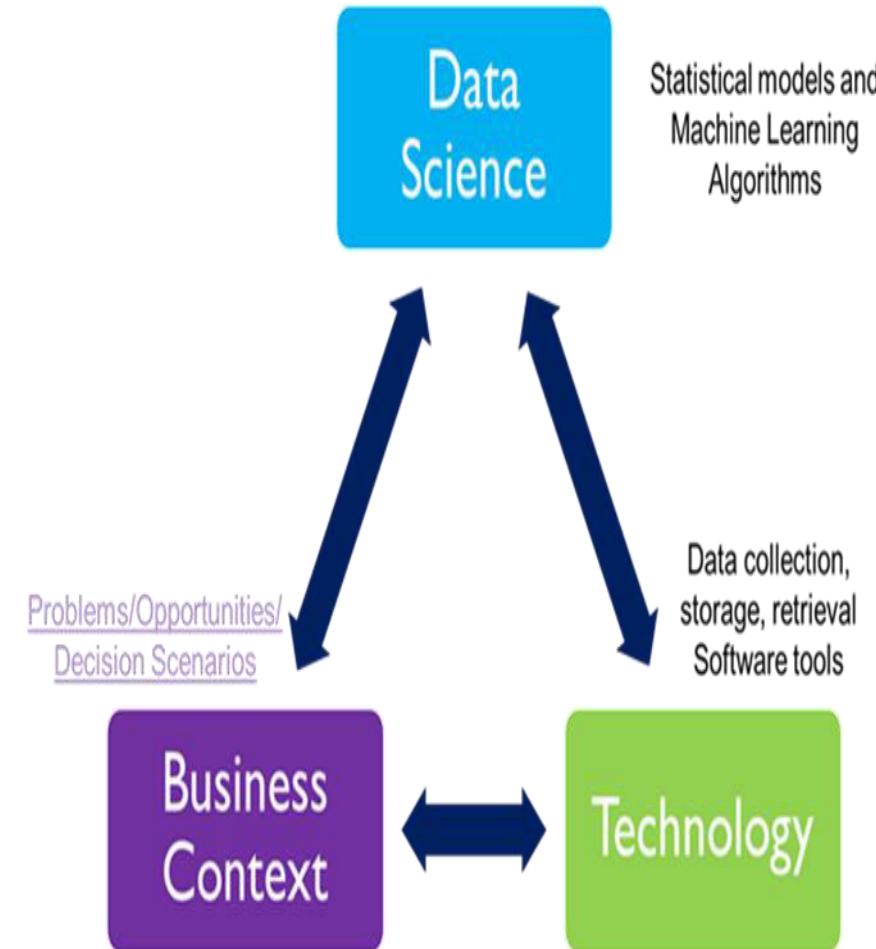
Problem context driving analytics : “Did you forget?”

feature

“Smart Basket” feature

The ability to predict the items that a customer may have forgotten to order can have a significant impact on the profits of online grocers such as bigbasket.com

The ability to ask right questions is an important success criteria for analytics projects.



Case Study

Indian online grocery store **bigbasket.com**

Technology:

To find out whether a customer has forgotten to place an order for an item

Information technology is used for data capture, data storage, data preparation, data analysis, data share and to deploy solution

An important output of analytics is automation of actionable items derived from analytical models which is usually achieved using IT

Case Study

Indian online grocery store bigbasket.com

Data science is the most important component of analytics, it consists of statistical and operations research techniques, machine learning and deep learning algorithms.

The objective of the data science component of analytics is to identify the most appropriate statistical model/machine learning algorithm that is best based on a measure of accuracy.

Example: “did you forget?” prediction is a classification problem in which customers are classified into:

1. Forget
2. Not forget

DATA ANALYTICS

Data Sources

- There has been enormous data growth in both commercial and scientific databases due to advances in data generation and collection technologies.
- New mantra
 - Gather whatever data you can whenever and wherever possible
- Expectations
 - Gathered data will have value either for the purpose collected or for a purpose not envisioned



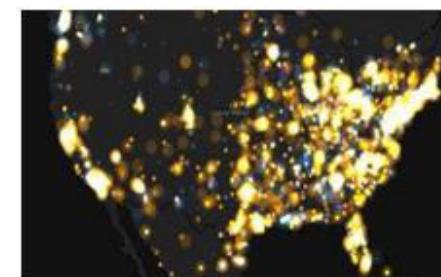
Cyber Security



E-Commerce



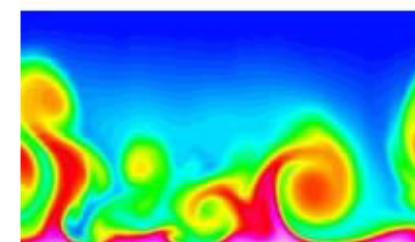
Traffic Patterns



Social Networking: Twitter



Sensor Networks



Computational Simulations

For example, data on customer purchasing habits might be collected to inform inventory decisions. But this same data could also be used to personalize marketing messages, predict future purchasing trends, or identify potential new products or services that customers might be interested in – these could be the purposes not initially envisioned.

Data Sources

- Lots of data is collected and warehoused every day
- Yahoo has peta bytes of web data
- Facebook has billions of active users
- Purchases at department/ grocery stores, e-commerce
 - Amazon handles millions of visits/day
- Bank/Credit Card transactions



How large is *big* (data)?

- 1 bit
- 1 byte = 8 bits
- 1 KB = 1024 bytes
- 1 MB = 1024 KB (kilobytes)
- 1 GB = 1024 MB (megabytes)
- 1 TB = 1024 GB (gigabytes) $\approx 10^{12}$ bytes
- 1 PB = 1024 TB (terabytes) $\approx 10^{15}$ bytes

20 PB = amt of data processed by Google per day!

- 1 EB = 1024 PB (petabytes)
- 1 ZB = 1024 EB (exabytes)
- 1 YB = 1024 ZB (zettabytes)
- What is a Domegemegrottebyte?

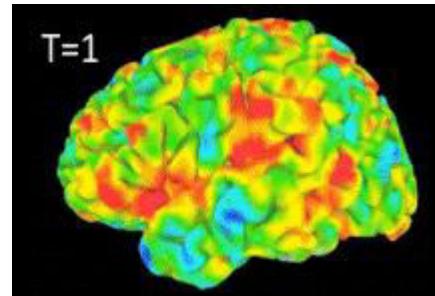
"Domegemegrottebyte" is a hypothetical unit of digital information storage, defined as 2^{128} bytes, or equivalent to 16^{32} bytes.

DATA ANALYTICS

Data Sources

Data collected and stored at enormous speeds

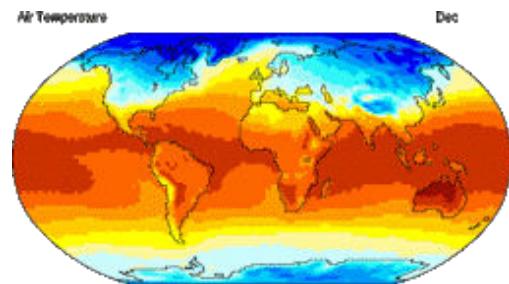
- Remote sensors on a satellite
 - NASA EOSDIS archives over petabytes of earth science data / year
- Telescopes scanning the skies
 - Sky survey data
- High-throughput biological data
- Scientific simulations - terabytes of data generated in a few hours



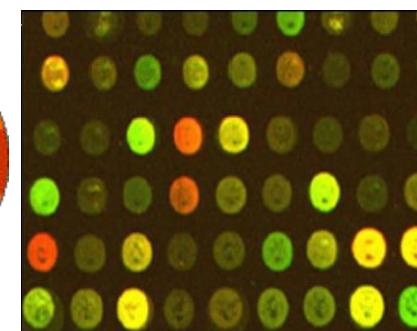
fMRI Data from Brain



Sky Survey Data



Surface Temperature of Earth



Gene Expression Data

DATA ANALYTICS

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**.

- An object is also known as a record, point, case, sample, entity, or instance.

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

- **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 values can be different

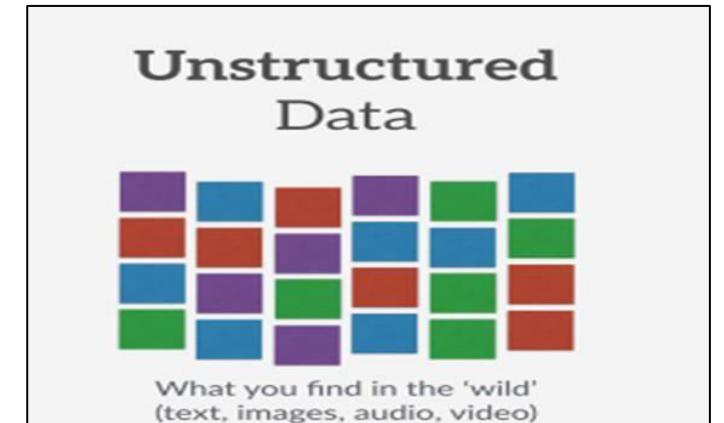
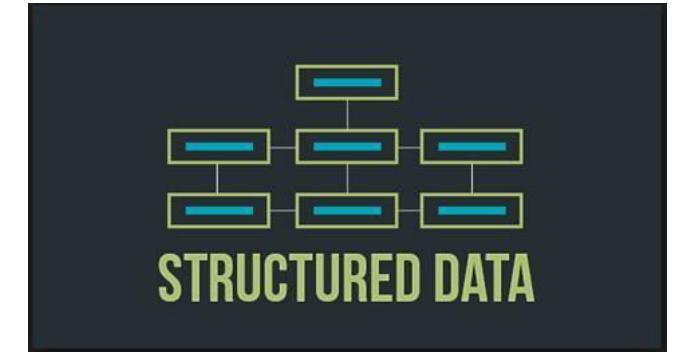
- There are different types of attributes
 - Nominal
 - Examples: ID numbers, eye color, zip codes
 - Ordinal
 - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height {tall, medium, short}
 - Interval
 - Examples: Calendar dates, temperatures in Celsius or Fahrenheit.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

Discrete and Continuous Attributes

- Discrete Attribute
 - Has only a finite or countably infinite set of values
 - Examples: zip codes, counts, or the set of words in a collection of documents
 - Often represented as **integer variables**.
 - Note: **binary attributes** are a special case of discrete attributes
- Continuous Attribute
 - Has real numbers as attribute values
 - Examples: 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**.

Data Representations

- **Structured Data:** Structured data means that the data is described in a matrix form with labelled rows and columns.eg: csv,rdbms,spreadsheets
- **Unstructured Data:** Any data that is not originally in the matrix form with rows and columns is an unstructured data.(tweets, videos,audios,scientific data)
- **Semi structured:** Semi-structured data (also known as partially structured data) is a type of data that doesn't follow the tabular structure associated with relational databases or other forms of data tables but does contain tags and metadata to separate semantic elements and establish hierarchies of records and fields.(XML,JSON). The key advantage of semi-structured data is its flexibility. It allows for variability in the data, accommodating unexpected attributes or nested structures, which can be especially useful when dealing with diverse data sources or evolving schemas.



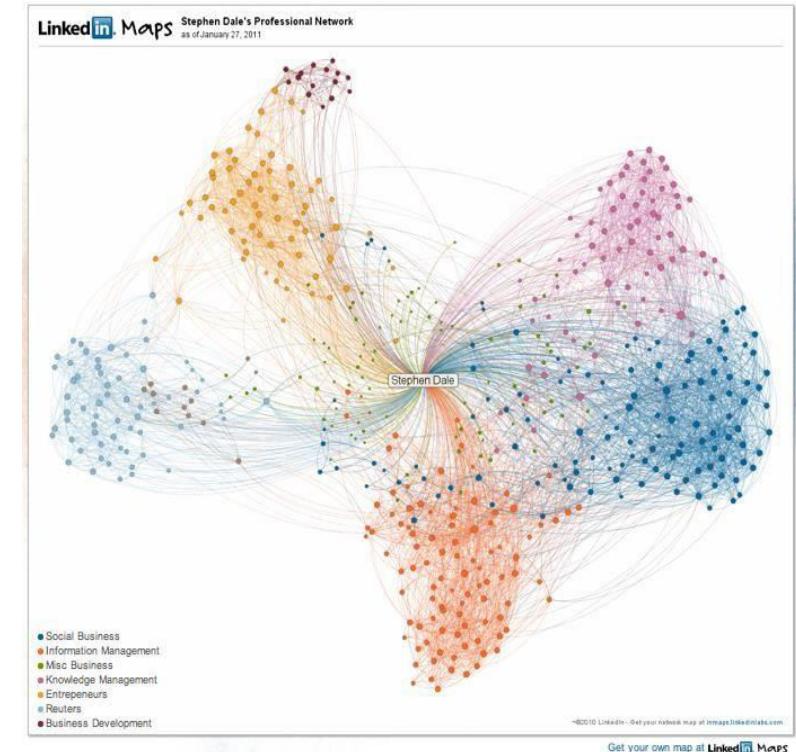
- Relational databases and spreadsheets. – **Structured Data**
- Text and multimedia content. Photos and graphic images, videos, streaming instrument data, webpages, PDF files, PowerPoint presentations, emails, blog entries, wikis and word processing documents. - **Unstructured Data**
- XML documents and NoSQL databases. – **Semi structured Data**
- For example, word processing software now can include metadata showing the author's name and the date created, with the bulk of the document just being unstructured text.

DATA ANALYTICS

Data Representations

- Record
 - Relational records
 - Data matrix, e.g., numerical matrix, crosstabs
 - Document data: text documents: term-frequency vector
 - Transaction data

- Graph and network
 - World Wide Web
 - Social or information networks
 - Molecular Structures



- Ordered
 - Video data: sequence of images
 - Temporal data: time-series
 - Sequential Data: transaction sequences
 - Genetic sequence data

- Spatial, image and multimedia
 - Spatial data: maps
 - Image data
 - Video data

Data Representations-Record Data

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

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

DATA ANALYTICS

Data Representations-Data Matrix

- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute
- Such data set can be represented by an m by n matrix, where there are m rows, one for each object, and n columns, one for each attribute

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

DATA ANALYTICS

Data Representations-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	w in	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Data Representations-Transaction data

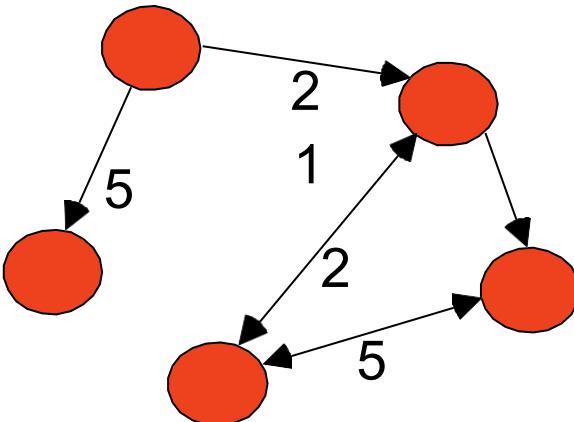
- A special type of record data, where
 - each record (transaction) involves a set of items.
 - For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.

<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

DATA ANALYTICS

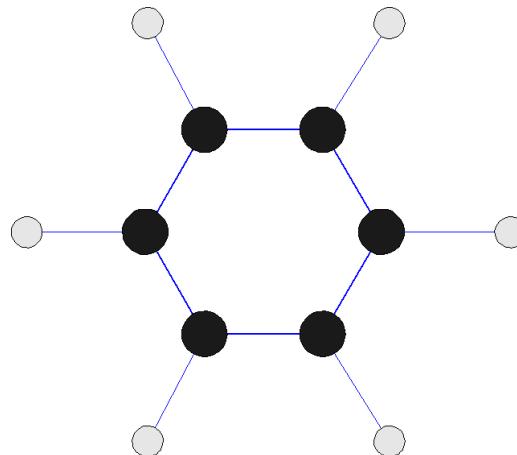
Data Representations

- Graph Data
- Examples: Generic graph and HTML Links



```
<a>Graph Partitioning </a>
<li>
<a href="papers/papers.html#aaaa">
Parallel Solution of Sparse Linear System of Equations </a>
</li>
<a href="papers/papers.html#fffff">
N-Body Computation and Dense Linear System Solvers
```

- Chemical Data
- Benzene Molecule: C₆H₆



DATA ANALYTICS

Data Representations-Ordered Data

- Sequences of transactions ■

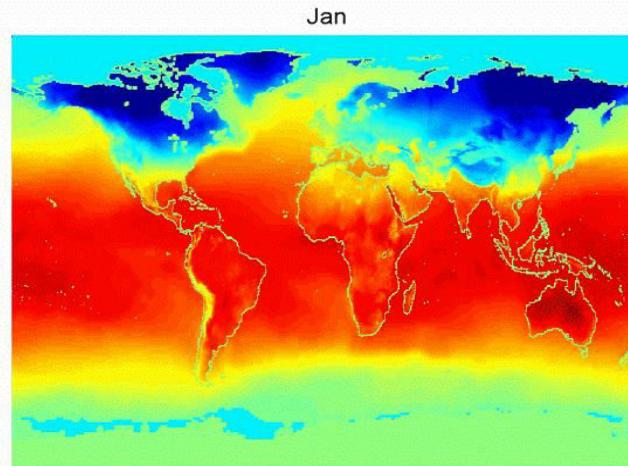
Items/Events

(A B) (D) (C E)
 (B D) (C) (E)
 (C D) (B) (A E)



**An element of
the sequence**

- Spatio-Temporal Data



- Genomic sequence data

```

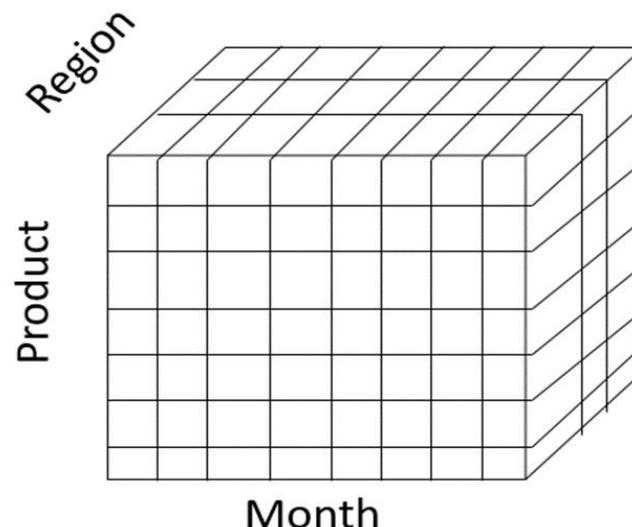
GGTTCCGCCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCGCGTC
GAGAAGGGCCCAGCCTGGCGGGCG
GGGGGAGGCAGGGCCGCCGAGC
CCAACCGAGTCCGACCAAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
    
```

DATA ANALYTICS

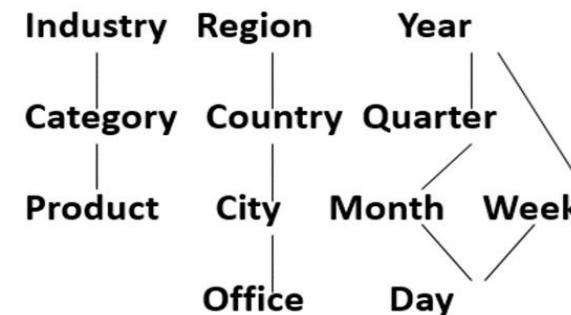
Data Representations- Data Warehouse

“A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management’s decision-making process.”—W. H. Inmon

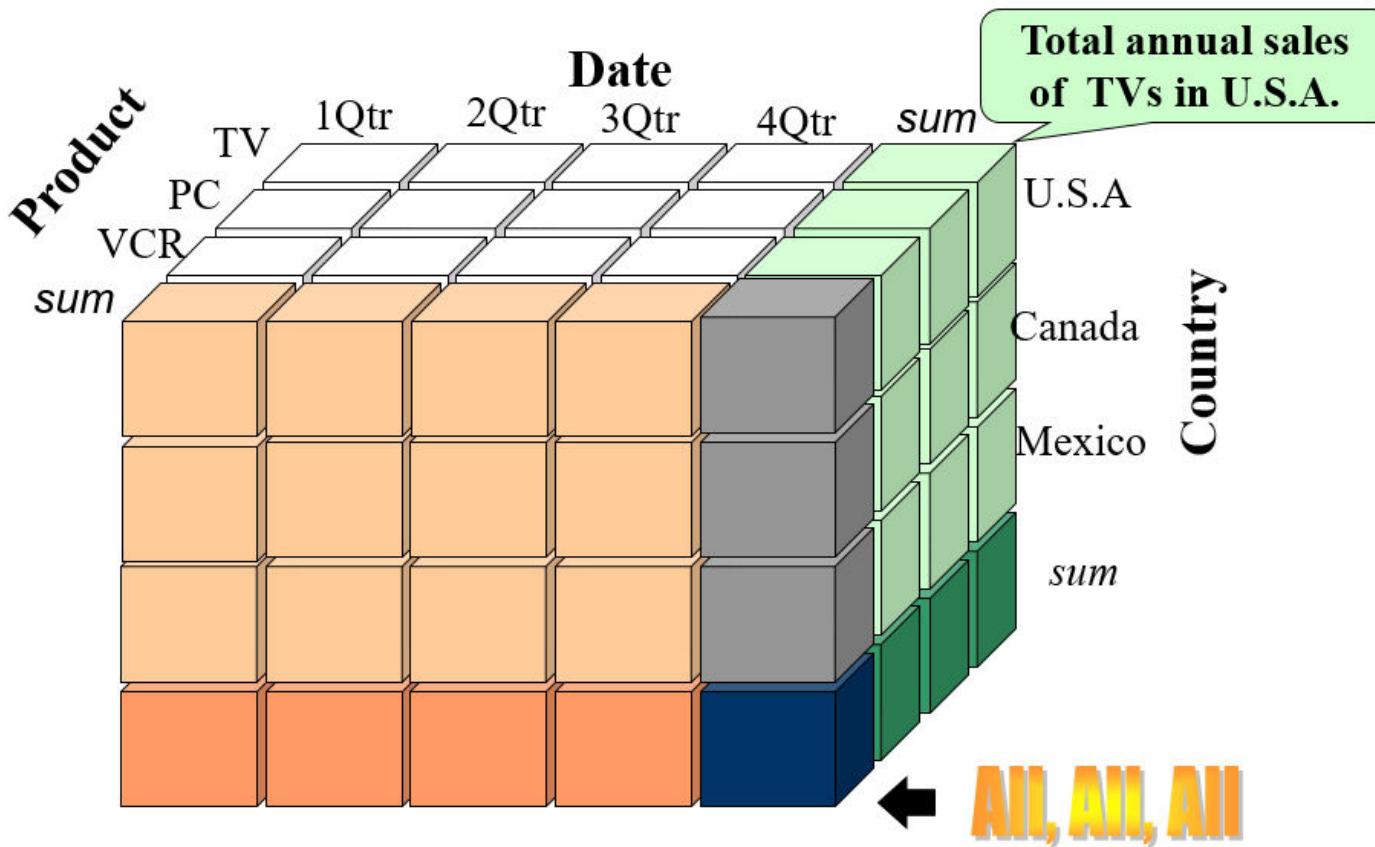
Sales volume as a function of product, month, and region



Dimensions: *Product, Location, Time*
Hierarchical summarization paths



A Sample Data Cube

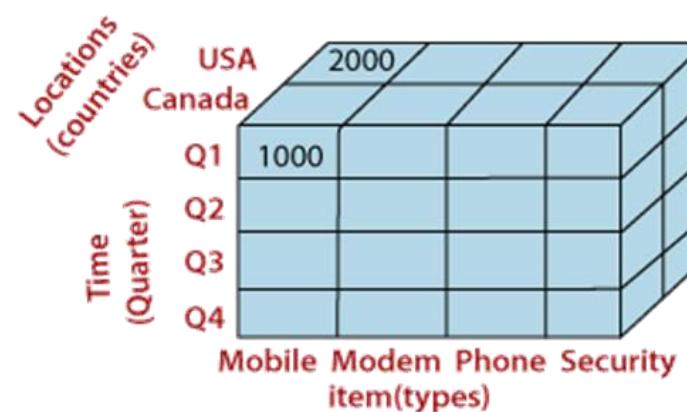
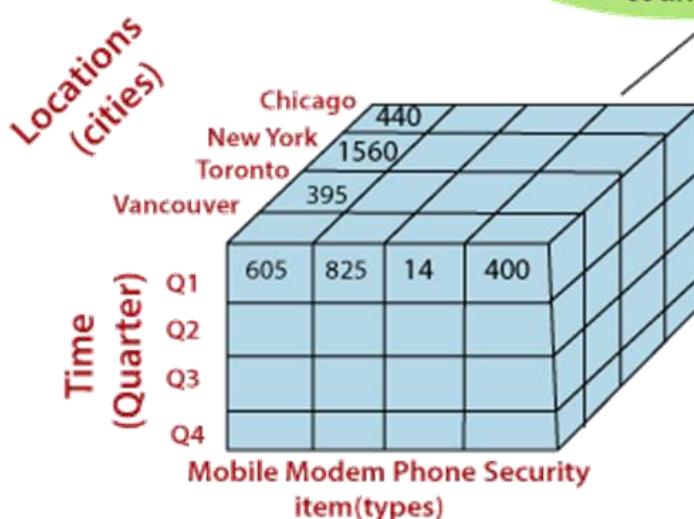


Typical OLAP Operations

- **Roll up (drill-up):** summarize data
 - by climbing up hierarchy or by dimension reduction
- **Drill down (roll down):** reverse of roll-up
 - from higher level summary to lower level summary or detailed data, or introducing new dimensions
- **Slice and dice:** project and select
- **Pivot (rotate):**
 - reorient the cube, visualization, 3D to series of 2D planes
- Other operations
 - **drill across:** involving (across) more than one fact table
 - **drill through:** through the bottom level of the cube to its back-end relational tables (using SQL)

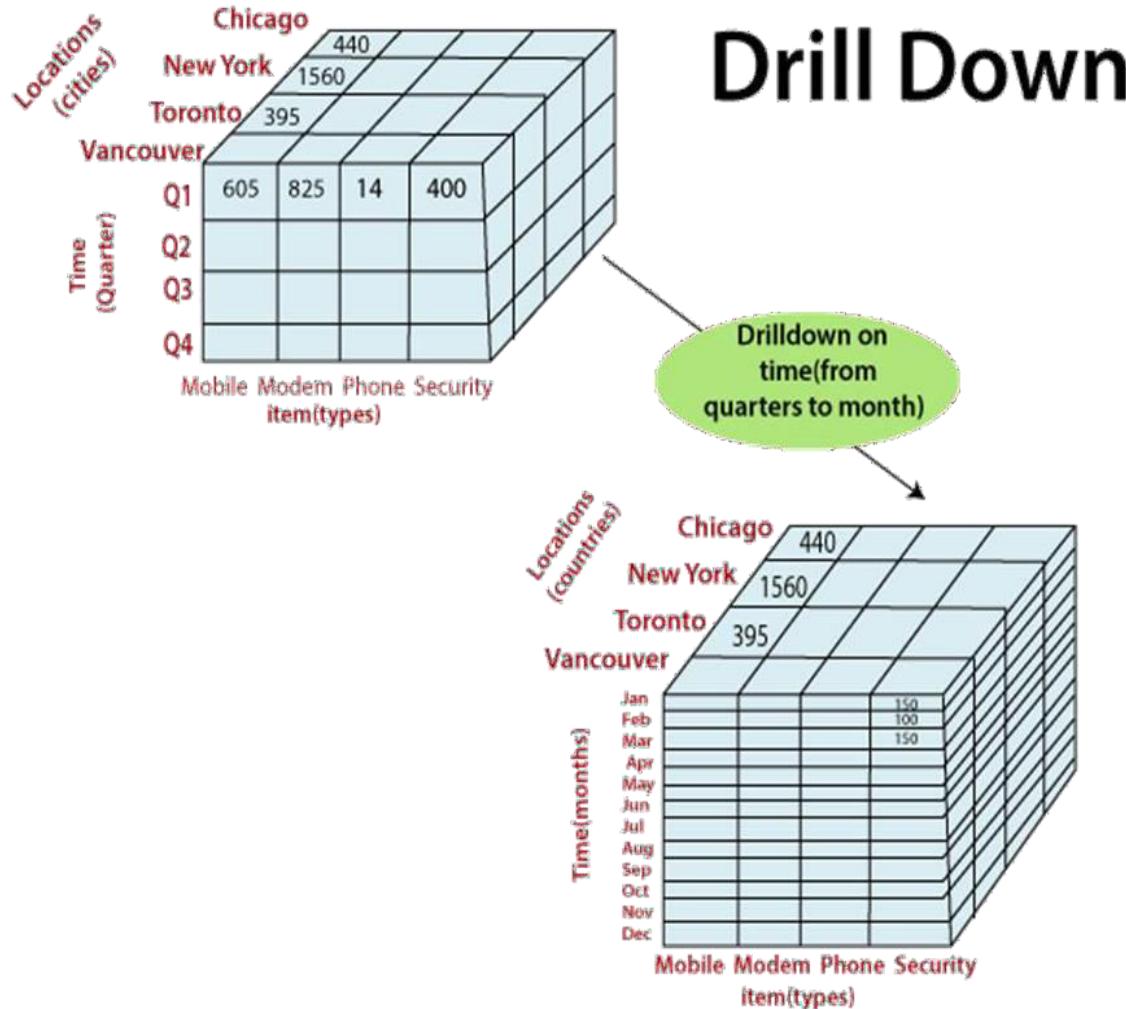
Typical OLAP Operations

Roll UP



Roll-up: operation and aggregate certain similar data attributes having the same dimension together. For example, if the data cube displays the daily income of a customer, we can use a roll-up operation to find the monthly income of his salary.
In our case we roll-up from cities to countries as shown

Typical OLAP Operations



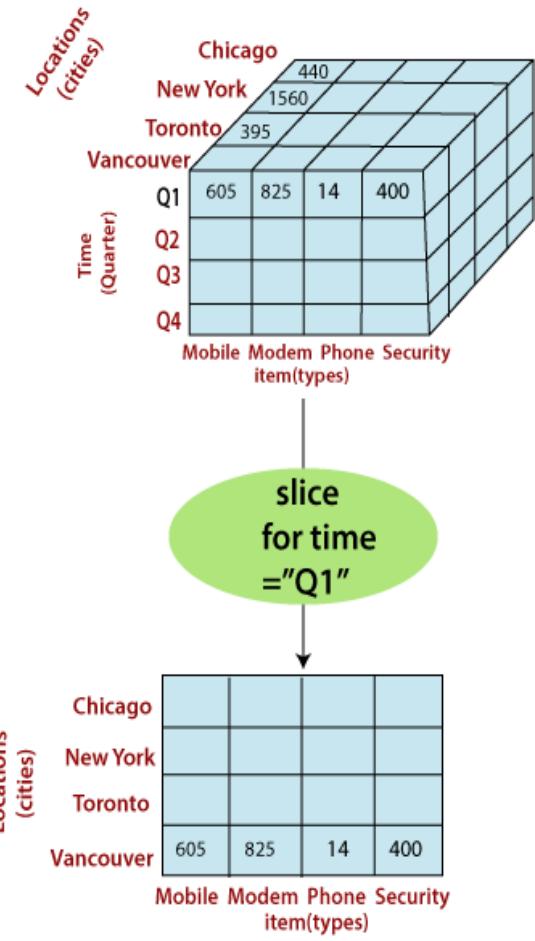
Drill Down

Drill-down: this operation is the reverse of the roll-up operation. It allows us to take particular information and then subdivide it further for coarser granularity analysis. It zooms into more detail. For example- if India is an attribute of a country column and we wish to see villages in India, then the drill-down operation splits India into states, districts, towns, cities, villages and then displays the required information.

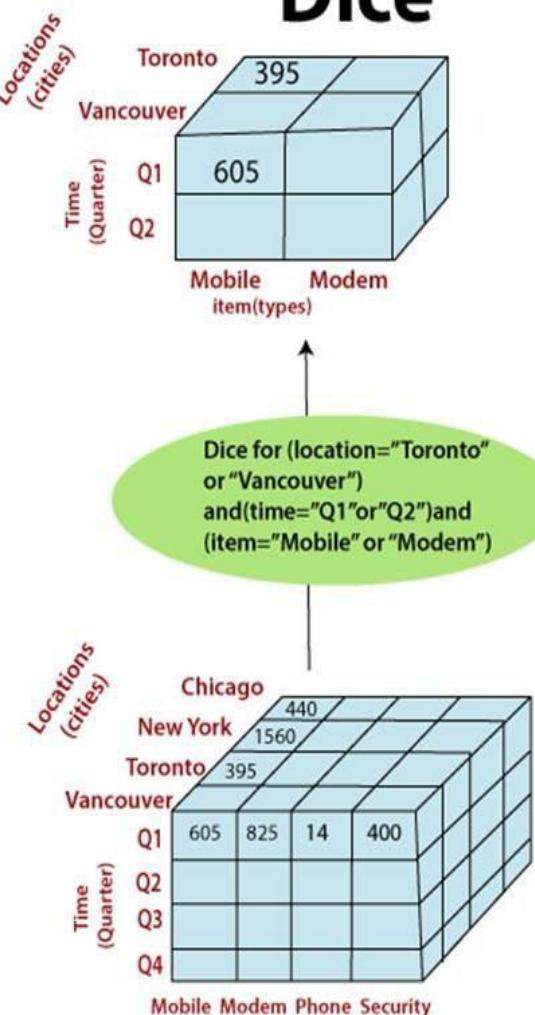
In our case we drill down from quarters to months.

Typical OLAP Operations

Slice



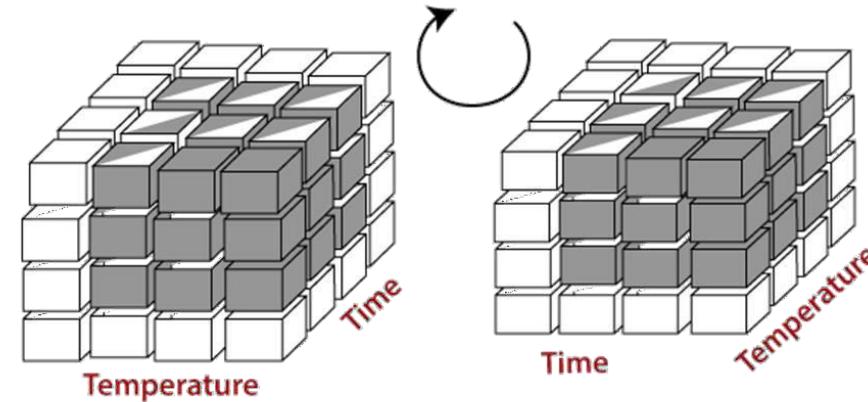
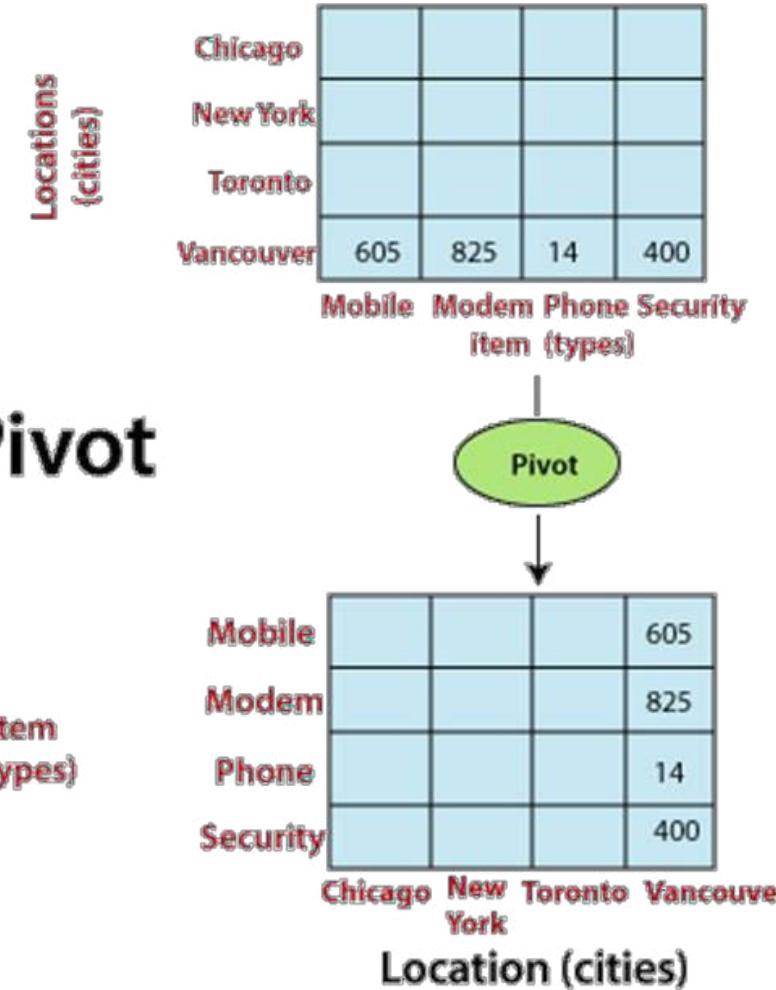
Dice



- **Slicing:** this operation filters the unnecessary portions. Suppose in a particular dimension, the user doesn't need everything for analysis, rather a particular attribute.

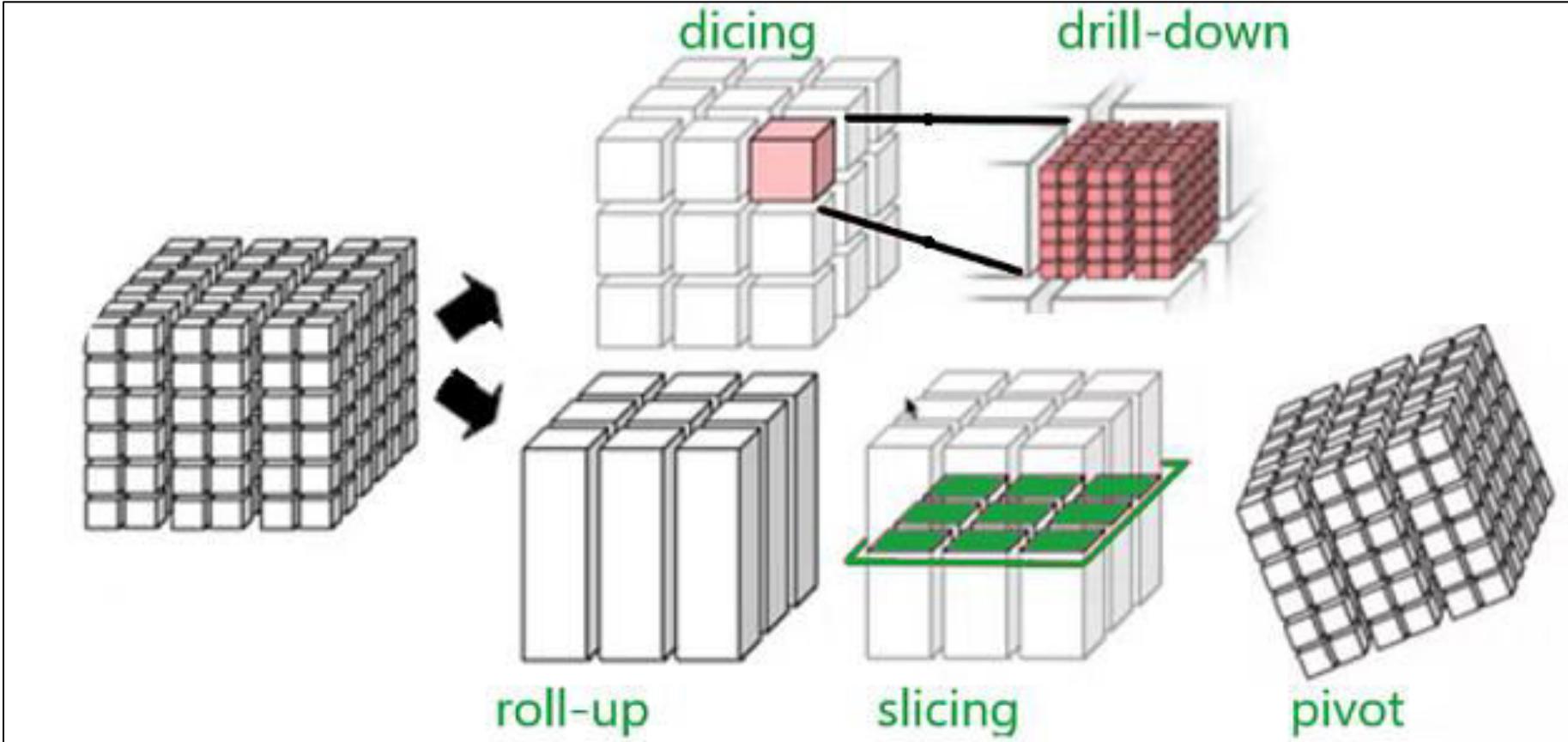
- **Dicing:** this operation does a multidimensional cutting, that not only cuts only one dimension but also can go to another dimension and cut a certain range of it. As a result, it looks more like a subcube out of the whole cube(as depicted in the figure).

Typical OLAP Operations



- **Pivot:** this operation is very important from a viewing point of view. It basically transforms the data cube in terms of view. It doesn't change the data present in the data cube. For example, if the user is comparing year versus branch, using the pivot operation, the user can change the viewpoint and now compare branch versus item type.

A Quick Glance on OLAP operations



Test your understanding

- To what type of an attribute does **shoe size** belong to?

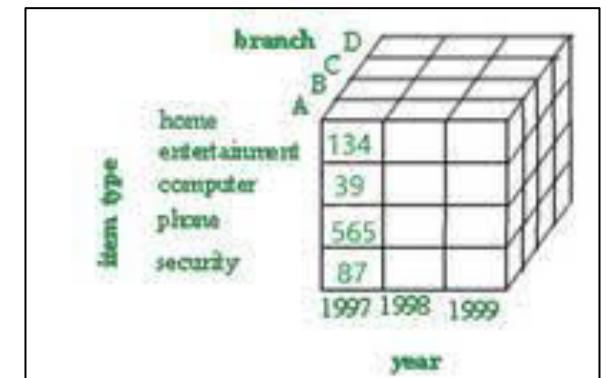
Interval

- Which OLAP operation are you likely to perform at the **end** of the financial year?

Roll-Up

- The example here is a 3D cube having attributes like branch(A,B,C,D),item type (home, entertainment, computer, phone,security),year(1997,1998,1999). If user wants to observe only “branch A” data then which OLAP operation must be performed?

Slicing



References

- Business Analytics by U. Dinesh Kumar – Wiley 2nd Edition, 2022
Chapter : 1.1-1.7
- Data Mining : Concepts and Techniques by Han, Kamber and Pei ,
The Morgan Kaufmann Series in Data Management Systems ,3rd
Edition Chapter : 4.2.5
- <https://www.geeksforgeeks.org/data-cube-or-olap-approach-in-data-mining/>



THANK YOU

Dr. Gowri Srinivasa

Professor, Department of Computer Science
and Engineering, PES University, Bengaluru

Email: gsrinivasa@pes.edu



DATA ANALYTICS

UE21CS342AA2

UNIT-1

**Lecture 2: Descriptive statistics
- a review**

Gowri Srinivasa

Department of Computer Science and Engineering

Data Analytics

Unit 1

Lecture 2 : Descriptive statistics

Slides excerpted from: U. Dinesh Kumar,
“Business Analytics”, Wiley, 2nd Edition 2022

Gowri Srinivasa
Department of Computer Science and Engineering

Slides collated by:

Nishanth M S PESU-2023, Department of CSE
nishanthmsathish.23@gmail.com

Harshitha Srikanth ,VII Sem ,PESU,Department of CSE
harshithasrikanth13@gmail.com

With grateful thanks for contribution of slides to:
Dr. Mamatha H R, Professor at the Department of CSE, PESU

Revisiting types of data attributes

Attribute Type	Description	Examples	Operations
Nominal	The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. ($=, \neq$)	zip codes, employee ID numbers, eye color, sex: { <i>male, female</i> }	mode, entropy, contingency correlation, χ^2 test
Ordinal	The values of an ordinal attribute provide enough information to order objects. ($<, >$)	hardness of minerals, { <i>good, better, best</i> }, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -)	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, electrical current	geometric mean, harmonic mean, percent variation

How is Interval different to Ratio?

Features	Interval scale	Ratio scale
Variable property	All variables measured in an interval scale can be added, subtracted, and multiplied. You cannot calculate a ratio between them.	Ratio scale has all the characteristics of an interval scale, in addition, to be able to calculate ratios. That is, you can leverage numbers on the scale against 0.
Absolute Point Zero	Zero-point in an interval scale is arbitrary. For example, the temperature can be below 0 degrees Celsius and into negative temperatures.	The ratio scale has an absolute zero or character of origin . Height and weight cannot be zero or below zero. (Zero means 'nothing'.)
Calculation	Statistically, in an interval scale, the arithmetic mean is calculated.	Statistically, in a ratio scale, the geometric or harmonic mean is calculated.
Measurement	Interval scale can measure size and magnitude as multiple factors of a defined unit.	Ratio scale can measure size and magnitude as a factor of one defined unit in terms of another.
Example	A classic example of an interval scale is the temperature in Celsius. The difference in temperature between 50 degrees and 60 degrees is 10 degrees; this is the same difference between 70 degrees and 80 degrees.	Classic examples of a ratio scale are any variable that possesses an absolute zero characteristic, like age, weight, height, or sales figures.

Transformations on Data

Attribute Level	Transformation	Comments
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., $\text{new_value} = f(\text{old_value})$ 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}.
Interval	$\text{new_value} = a * \text{old_value} + b$ where a and b are constants $F = 9/5 \times C + 32$	Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree).
Ratio	$\text{new_value} = a * \text{old_value}$	Length can be measured in meters or feet.

Test your understanding!

Classify the following data :

- Time of the day in terms of AM/PM

Qualitative ,Nominal ,Discrete

- Angles measured in degrees between 0 and 360

Quantitative ,Ratio ,Continuous

- Medals awarded at the Olympics

Qualitative ,Ordinal ,Discrete

- Brightness as measured by a light meter

Quantitative, Ratio ,Continuous

- Brightness as measured by people's judgements

Qualitative, Ordinal ,Discrete

Classification based on nature of data collection

- **Cross-Sectional Data** : Data collected on many variables of interest at the same instance or duration of time. Example : Data of all sitcoms released in 2022. The attributes can be budget, actors , popularity , social media engagement and so on.
- **Time-series Data** : Data collected on a single variable of interest over several time intervals , like on a daily , monthly or a weekly basis. Example : Daily price of Bitcoin since its inception.
- **Panel Data** : Data collected on several variables (multiple dimensions) over several time intervals. It is also known as **longitudinal data**. Example : Data collected on GDP (Gross Domestic Product) , Gini Index and rate of unemployment for several countries over several years.

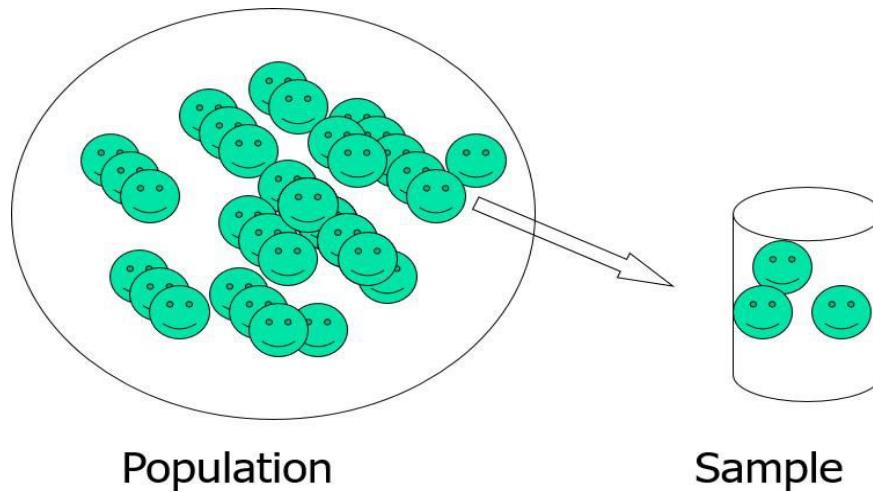
Identify which data?

- the gross annual income for each of 1000 randomly chosen households in New York City for the year 2000.
- weekly sales graph of an ice-cream sold during a holiday period at some shop.

- Cross sectional data means that we have data from many units, at one point in time.
- Time series data means that we have data from one unit, over many points in time.
- Panel data (or time series cross section) means that we have data from many units, over many points in time.

Population and Sample

- **Population** : The set of all possible observations or records (also called records , cases, subjects or data points) for a given context of the problem.
- **Sample** : The subset taken from the population. In real world scenarios , an inference is made about the population based on sample data.



Summary statistics

- **Exploratory data analysis (EDA)** : A preliminary exploration of data to better understand its characteristics.

Key motivations :

- Helps in selecting the right tool for preprocessing or analysis in the further stages.
 - Makes use of humans' abilities to recognize patterns ,some of which aren't covered by data analysis tools.
-
- **Summary statistics** : Numbers that summarize properties of data. It is a part of EDA. Most of them can be calculated in a single pass of the data. Summarized properties include frequency , location and spread of the data.

Example : location – mean , spread – standard deviation

An illustration : Which group is smarter?

Consider the IQ scores of 13 students of two classes

Class A	Class B
102	127
128	131
131	96
98	80
140	93
93	120
110	109
115	162
109	103
89	111
106	109
119	87
97	105

Each individual may be different. If you try to understand a group by remembering the qualities of each member, you become overwhelmed and fail to understand the group.

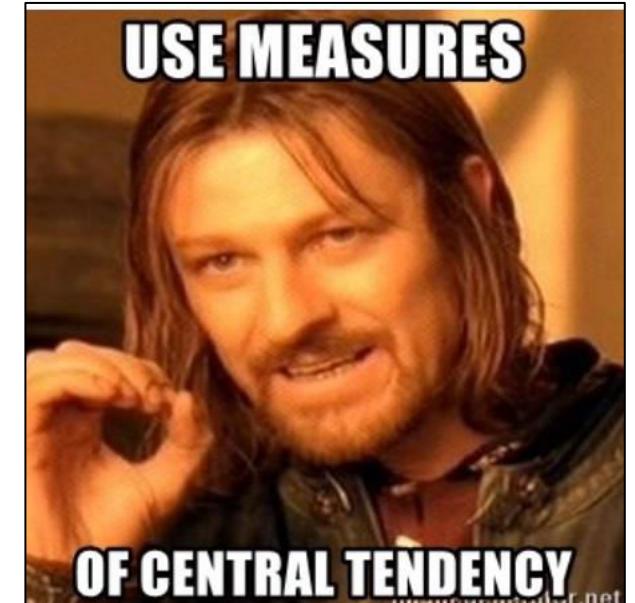
An illustration : Which group is smarter?

Class A – Average IQ	Class B – Average IQ
110.54	110.23

With this data , we can conclude that both classes are equally smart as their average IQs are roughly the same.

The question is easily answered now thanks to a descriptive summary statistic.

- Measures of central tendency are those used to describe the data using a single value.
- The three most frequently used measures of central tendencies are:
 - Mean
 - Median
 - Mode
- It helps in comparisons between different datasets.
- It helps in summarizing and comprehending the data.



- **Mean (or Average) Value**

Mean is the arithmetical average value of the data and is one of the most used measures of average tendency.

If n is the number of records in the sample and x_i is the value of i^{th} record, then the mean value is given by :

$$\text{Mean} = \bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \sum_{i=1}^n \frac{x_i}{n}$$

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

If the data is captured in frequencies, then use:

1) Find the mean for the following data

The frequency of age of students in Table 2.1 is given below:

Age	21	22	23	24
Frequency	1	4	3	2

2) Find the median for the following data

Consider the example of a bank. The number of deposits in a branch of a bank in a week is shown in Table 2.3.

TABLE 2.3 Number of daily deposits in a Bank

Day	1	2	3	4	5	6	7
Number of Deposits	245	326	180	226	445	319	260

Another example is the salary in Table 2.1 that can be arranged as follows:

180000, 220000, 235000, 240000, 240000, 240000, 250000, 270000, 300000, 425000

Mean

Symbol \bar{X} is frequently used to represent the estimated value of the mean from a sample. If the entire population is available and if we calculate mean based on the entire population, then we have the population mean which is denoted by μ (population mean).

Property of Mean

An important property of mean is that the summation of deviation of observations from the mean is zero, that is

$$\sum_{i=1}^n (X_i - \bar{X}) = 0$$

One should be careful about taking decisions based on the mean value of the data. There is a famous joke in statistics which says that "if someone's head is in freezer and leg is in the oven, the average body temperature would be fine, but the person may not be alive". Making decisions solely based on mean value is not advisable.

Median

- **Median (or Mid) Value**

Median is the value that divides the data into two equal parts. That is , the proportion of records below and above the median will be 50% each.

- **Finding the median**

- Arrange the data in increasing order.
- If number of observations n is odd , median is the observation at the position $(n+1)/2$.
- If n is even , the median is the average value of observations at positions $(n/2)$ and $(n+2)/2$.

Median is not calculated using the entire dataset like mean. We are simply looking for the midpoint rather than using the values of the entire data.

However , it is more stable than mean as adding a new observation doesn't change the median significantly.

Mode

- **Mode** is the most frequently occurring value in the dataset.
- It is the only measure of central tendency which is applicable to qualitative (nominal) data , since mean and median for nominal data are meaningless.
- For example , assume that there is student data with students' mode of transport , namely car , bus , two wheeler or the metro. Mean and Median are meaningless to analyze nominal data such as mode of transport.
- It is possible for a dataset to not have a mode at all. It occurs when **each** value of the dataset appears **equal** number of times.

Test your understanding!

- Which central tendency (when applicable and exists) provides a value which is present in the dataset?

Mode

- Which central tendency is not robust to outliers?

Mean

- When the dataset contains outliers, which measure of central tendency is the most appropriate to use?

Median

- Which function in R is used to load a package?

library()

Which of the following variable is an interval scale variable?

- (a) Age (b) Latitude and longitude (c) Marital status (d) Hair colour

References

- Business Analytics by U. Dinesh Kumar – Wiley 2nd Edition, 2022
Chapter 2.1-2.5
- Introduction to Data Mining , Tan, Steinbach, Kumar, 2nd Edition
- <https://www.tutorialspoint.com/>



THANK YOU

Dr. Gowri Srinivasa

Professor, Department of Computer Science
and Engineering, PES University, Bengaluru

Email: gsrinivasa@pes.edu



DATA ANALYTICS

UE21CS342AA2

UNIT-1

Lecture 3 : Descriptive Statistics - 2

Gowri Srinivasa

Department of Computer Science and Engineering

Data Analytics

Unit 1

Lecture 3 : Descriptive Statistics - 2

Slides excerpted from: U. Dinesh Kumar,
“Business Analytics”, Wiley, 2nd Edition 2022

Gowri Srinivasa

Department of Computer Science and Engineering

Slides collated by:

Nishanth M S PESU-2023, Department of CSE
nishanthmsathish.23@gmail.com

Harshitha Srikanth ,VII Sem ,PESU, Department of CSE
harshithasrikanth13@gmail.com

With grateful thanks for contribution of slides to:
Dr. Mamatha H R, Professor at the Department of CSE, PESU

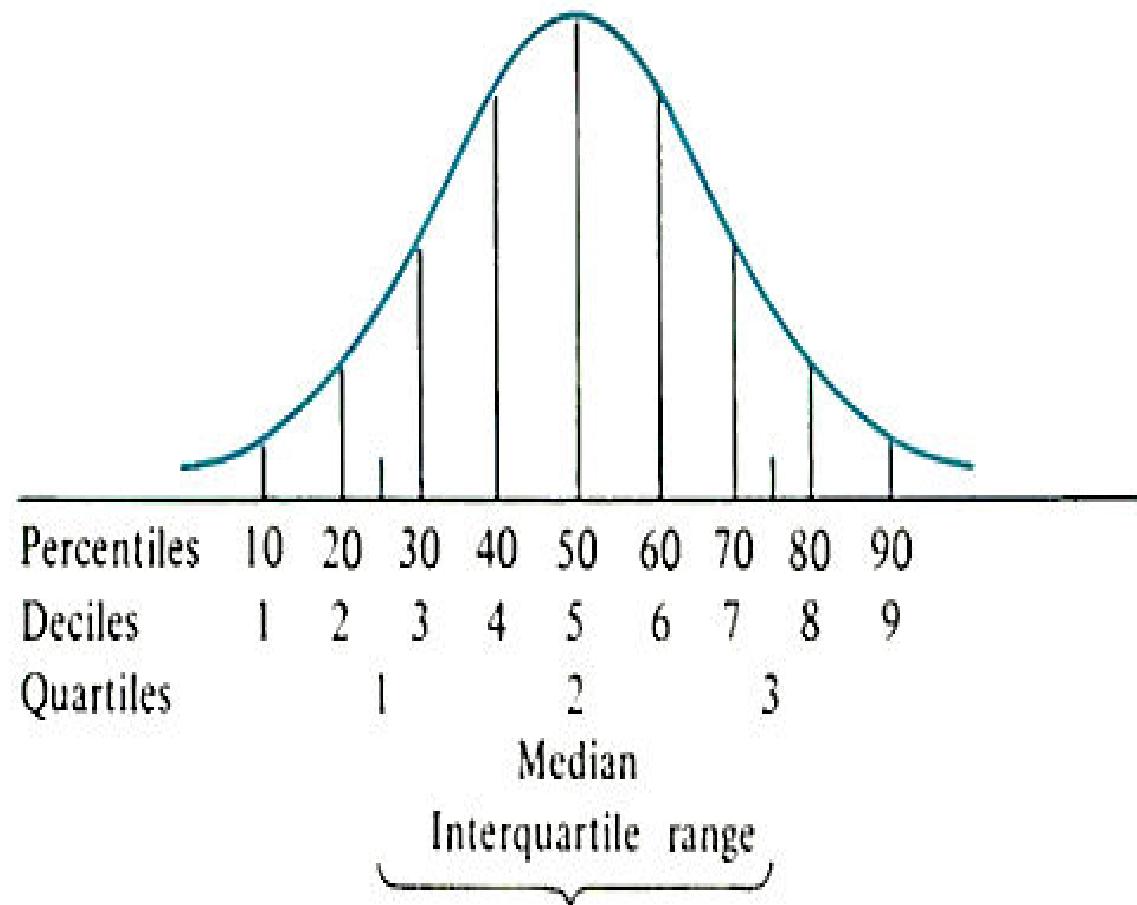
Percentile

- **Percentile**, denoted as P_x , is the value of the data at which x percentage of the data lies below that value. **It is used to identify the position of the observation in the data set.**
- For example , P_{10} denotes the value below which 10% of the data lies.
- In the context of asset management and reliability , P_{10} life implies the time by which 10% of the products fail.
- **To find P_x , the data must be arranged in ascending order.**

Position corresponding to $P_x \approx x(n+1)/100$

where P_x is the position in the data calculated and n is the number of

- **Decile** corresponds to special values of percentile that divide the data into 10 equal parts. First decile contains first 10% of the data and second decile contains first 20% of the data and so on.
- **Quartile** divides the data into 4 equal parts. The first quartile (Q_1) contains first 25% of the data, Q_2 contains 50% of the data and is also the median. Quartile 3 (Q_3) accounts for 75% of the data.



Example

Time between failures of wire-cut (in hours)

2	22	32	39	46	56	76	79	88	93
3	24	33	44	46	66	77	79	89	99
5	24	34	45	47	67	77	86	89	99
9	26	37	45	55	67	78	86	89	99
21	31	39	46	56	75	78	87	90	102

1. Calculate the mean, median, and mode of time between failures of wire-cuts
2. The company would like to know by what time 10% (ten percentile or P_{10}) and 90% (ninety percentile or P_{90}) of the wire-cuts will fail?
3. Calculate the values of P_{25} and P_{75} .

Solution

1. Mean = 57.64, median = 56, and mode = 46

1. Note that the data in the table is arranged in increasing order of columns. The position of $P_{10} = 10 \times (51)/100 = 5.1$. We can round off 5.1 to its nearest integer 5. The corresponding value from the table is 21. (10 % of the observations have a value of less than or equal to 21. That is by 21 hours, 10% of the wire-cuts will fail.

Instead of rounding the value obtained from the equation , we can use the

following approximation: Value at 5th position is 21. Value at position 5.1 is

approximated as $21 + (0.1 * (\text{value at } 6^{\text{th}} \text{ position} - \text{value at } 5^{\text{th}} \text{ position}))$
 $= 21 + (0.1 * 1) = 25.1$

Solution

$$P_{90} = 90 \times 51/100 = 45.9$$

The value at position 45 is 90, position 46 is 93. The value of position 45.9 is $90 + (0.9 * 3) = 92.7$. That is , 90% of the wire-cuts will fail by 92.7 hours.

3. $P_{25} = (1^{\text{st}} \text{ Quartile or } Q_1) = 25 \times 51/100 = 12.75$

Value at 12th position is 33, so

$$\begin{aligned} P_{25} &= 33 + 0.75 * (\text{Value at } 13^{\text{th}} \text{ position} - \text{Value at } 12^{\text{th}} \text{ position}) \\ &= 33 + 0.75 * 1 = 33.75 \end{aligned}$$

$$P_{75} = (3^{\text{rd}} \text{ Quartile or } Q_3) = 75 \times 51/100 = 38.25$$

Value at 38th position is 86, so

$$\begin{aligned} P_{75} &= 86 + 0.25 * (\text{Value at } 39^{\text{th}} \text{ position} - \text{Value at } 38^{\text{th}} \text{ position}) \\ &= 86 + 0.25 * 0 = 86 \end{aligned}$$

DATA ANALYTICS

Measures of Variation(**amount of dispersion in a dataset. How spread out are the values**)



- One of the primary objectives of analytics is to understand the **variability in the data and the cause of such variability**.
- Predictive analytics techniques such as regression attempt to explain variation in the outcome variable (Y) using predictor variables (X)
- Measures of variability are useful in identifying how close records are to the mean value and outliers in the data.
- An important application of variability is in feature selection during model building. If a variable has very **low variability**, it is unlikely to have a statistically significant relationship with the outcome variable.
- Variability in the data is measured by **range , Inter-Quartile distance (IQD), Variance , Standard Deviation and coefficient of variation**.

low variability

If a variable has very low variability, it essentially means that its values do not change much and are almost constant. Therefore, it brings very little information to the model, because it's not showing any notable difference across observations

Range , IQD and Variance

- **Range** is the difference between maximum and minimum value of the data. It captures the data spread.
- **Inter-quartile distance (IQD)**, also called inter-quartile range (IQR) is a measure of the distance between Quartile 1 (Q_1) and Quartile 3 (Q_3).

IQD is used in identifying outliers in the data. Values of data below $Q_1 - 1.5 \text{ IQD}$ and above $Q_3 + 1.5 \text{ IQD}$ are classified as potential outliers.

- **Variance** is a measure of variability in the data from the **mean value**.
Variance for population, σ^2 , is calculated using

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

makes our measure more sensitive to outliers, which may be desirable in some contexts.
Also get rid of negative sign.

Note that, in this equation, deviation from mean is squared since sum of deviations from mean will always add up to zero.

Sample Variance

- In case of a sample , the Sample Variance (S^2) is calculated using

$$S^2 = \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n-1}$$

- While calculating sample variance S^2 , the sum of squared deviation is divided by $(n-1)$, this is known as Bessel's correction.

Bessel's correction is a crucial adjustment when dealing with sample data. It highlights a fundamental concept in statistics: estimators derived from samples may have inherent biases, and sometimes corrections are necessary to obtain more accurate and unbiased estimations for population parameters.

Standard Deviation

- **Standard Deviation:** It's the square root of the variance. For a population:

$$\sigma = \sqrt{\sigma^2}$$

And for a sample:

$$s = \sqrt{s^2}$$

- The population standard deviation (σ) and sample standard deviation (S) are given by

$$\sigma = \sqrt{\sum_{i=1}^n \frac{(X_i - \mu)^2}{n}}$$

$$S = \sqrt{\sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n-1}}$$

- Standard deviation is preferred to variance as it is in the same unit as the original values and the mean , which makes it easier to interpret and understand the variability in the data and its deviation from the mean.

TABLE 2.6 Underestimation of standard deviation in sample

Data	Standard deviation (using sample mean 53.2)	Standard deviation (using population mean 57.64)
2	2621.44	3095.81
3	2520.04	2985.53
5	2323.24	2770.97
9	1953.64	2365.85
21	1036.84	1342.49
93	1584.04	1250.33
99	2097.64	1710.65
99	2097.64	1710.65
99	2097.64	1710.65
102	2381.44	1967.81
Sample Mean = 53.2	$\sum (X_i - \bar{X})^2 = 20713.60$	$\sum (X_i - \mu)^2 = 20910.74$

In Table 2.6, we can see that the numerator in Eq. (2.4) is underestimated (20713.60) when we use the sample average against population average (20910.74). This will result in underestimation of the standard deviation, a phenomenon called **downward bias**. To overcome this bias, we divide $\sum (X_i - \bar{X})^2$ with $(n - 1)$ instead of n .

Degrees of Freedom

There are two definitions to interpret degrees of freedom :

- **Degrees of freedom** is equal to the number of independent variables in the model. For example, we can create any sample of size n with mean value of X by randomly selecting $(n-1)$ values. We need to fix just one out of n values. Thus the number of independent variables in this case is $(n-1)$.
- **Degrees of freedom** is defined as the difference between the number of observations in the sample and number of parameters estimated. If there are n observations in the sample and k parameters are estimated , then the degrees of freedom is $(n-k)$.

Whenever we estimate a parameter from a sample, we lose a degree of freedom.

Chebyshev's Theorem

- **Chebyshev's theorem** (also known as Chebyshev's inequality) is an empirical rule that allows us to predict the proportion of observations that is likely to lie between an interval defined using mean and standard deviation.
- Probability of finding a randomly selected value in an interval defined by $\mu \pm k\sigma$ is $1 - \frac{1}{k^2}$. That is ,

$$P(\mu - k\sigma \leq X \leq \mu + k\sigma) \geq 1 - \frac{1}{k^2}$$

- Alternatively , it can be written as

$$\Pr(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

In Chebyshev's Theorem, 'k' is a constant that stands for the number of standard deviations from the mean. It is a positive number ($k > 1$).

Example

Amount spent per month by a segment of credit card users of a bank has a mean value of 12000 and a standard deviation of 2000. Calculate the proportion of customers who are spending between 8000 and 16000.

- Solution:

$$\begin{aligned} & P(8000 \leq X \leq 16000) \\ & = P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \geq 1 - \frac{1}{2^2} = 0.75 \end{aligned}$$

That is, the proportion of customers spending between 8000 and 16000 is at least 0.75 (or 75%).

Chebyshev's Theorem

So now let's look at an example. Suppose 1,000 applicants show up for a job interview, but there are only 70 positions available. To select the best 70 people amongst the 1,000 applicants, the employer gives an aptitude test to judge their abilities. The mean score on the test is 60, with a standard deviation of 6. If an applicant scores an 84, can they assume they are getting a job?

$$\text{If } \mu = 60 \quad \sigma = 6 \quad k = ? \quad P(x > 84) = ?$$

$$|X - \mu| \rightarrow |84 - 60| = 24 \quad \text{and} \quad 24 = \underbrace{4(6)}_{k\sigma} \quad \text{so} \quad k = 4$$

$$\text{If } P(|X - \mu| > k\sigma) \leq \frac{1}{k^2}, \text{ then } P(X > 84) \leq \frac{1}{(4)^2} = 0.0625$$

$$1,000(0.0625) = 62.5$$

Measures of Shape – Skewness and Kurtosis

- **Skewness** is a measure of symmetry or lack of symmetry. A dataset is symmetrical when the proportion of data at equal distance (measured in terms of standard deviation) from mean (or median) is equal. That is, the proportion of data between μ and $\mu - k\sigma$ is same as μ and $\mu + k\sigma$, where k is some positive constant.
- **Pearson's moment coefficient of skewness for a dataset with n observations is given by**

$$g_1 = \frac{\sum_{i=1}^n (X_i - \mu)^3 / n}{\sigma^3}$$

- The value of g_1 will be close to 0 when the data is symmetrical. A positive value of g_1 indicates a positive skewness and a negative value indicates negative skewness.

Skewness

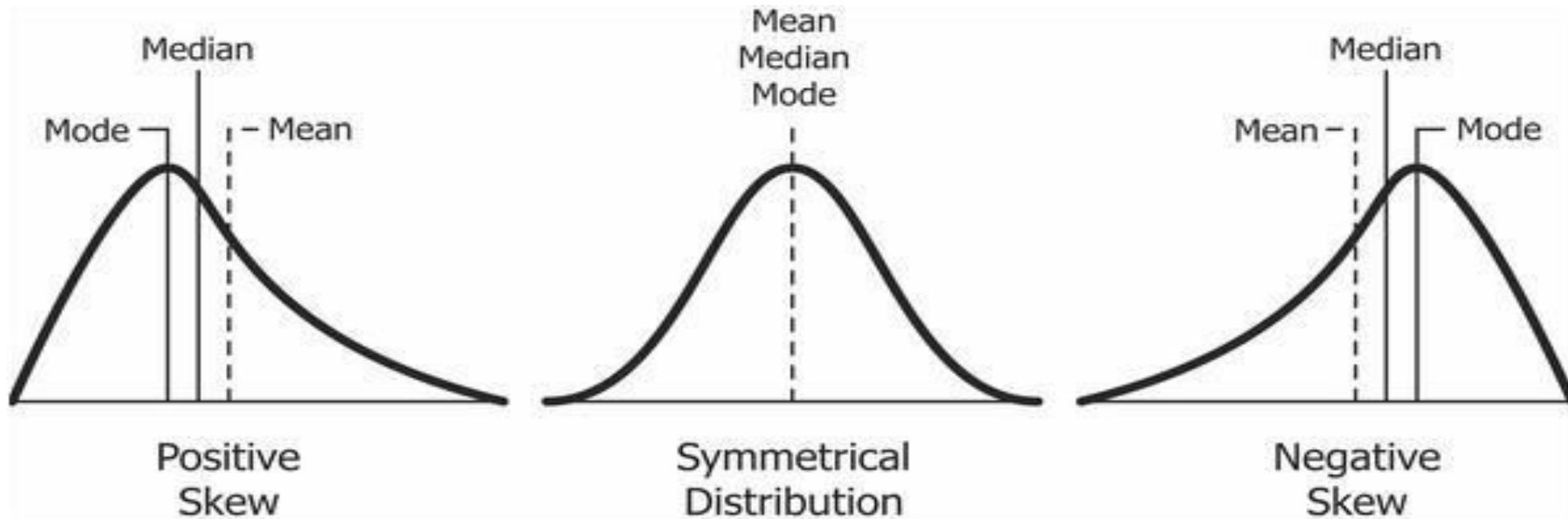
- The following formula is used usually for a sample with n observations (Joanes and Gill, 1998):

$$G_1 = \frac{\sqrt{n(n-1)}}{n-2} g_1$$

- The value of $\frac{\sqrt{n(n-1)}}{n-2}$ will converge to 1 as the value of n increases.
(If the value of G_1 is negative, we can conclude that the data is left skewed)
- Skewness in finance is used to understand risk and return. For example, negative skewness about data on return on stocks would imply that the returns could be much lower than the mean, and it could result in a loss. Positive skewed distribution of returns would imply the returns could be much higher than average.

Symmetric vs Skewed Data

- Median, mean and mode of symmetric, positively skewed(right tailed) and negatively skewed(left tailed)



Kurtosis

- Kurtosis is another measure of shape, aimed at the shape of the tail, that is, whether the tail of the data distribution is heavy or light.

$$\text{Kurtosis} = \frac{\sum_{i=1}^n (X_i - \bar{X})^4 / n}{\sigma^4}$$

- Kurtosis value of less than 3 is called platykurtic distribution and greater than 3 is called leptokurtic distribution. Kurtosis value of 3 indicates standard normal distribution and is called as mesokurtic.

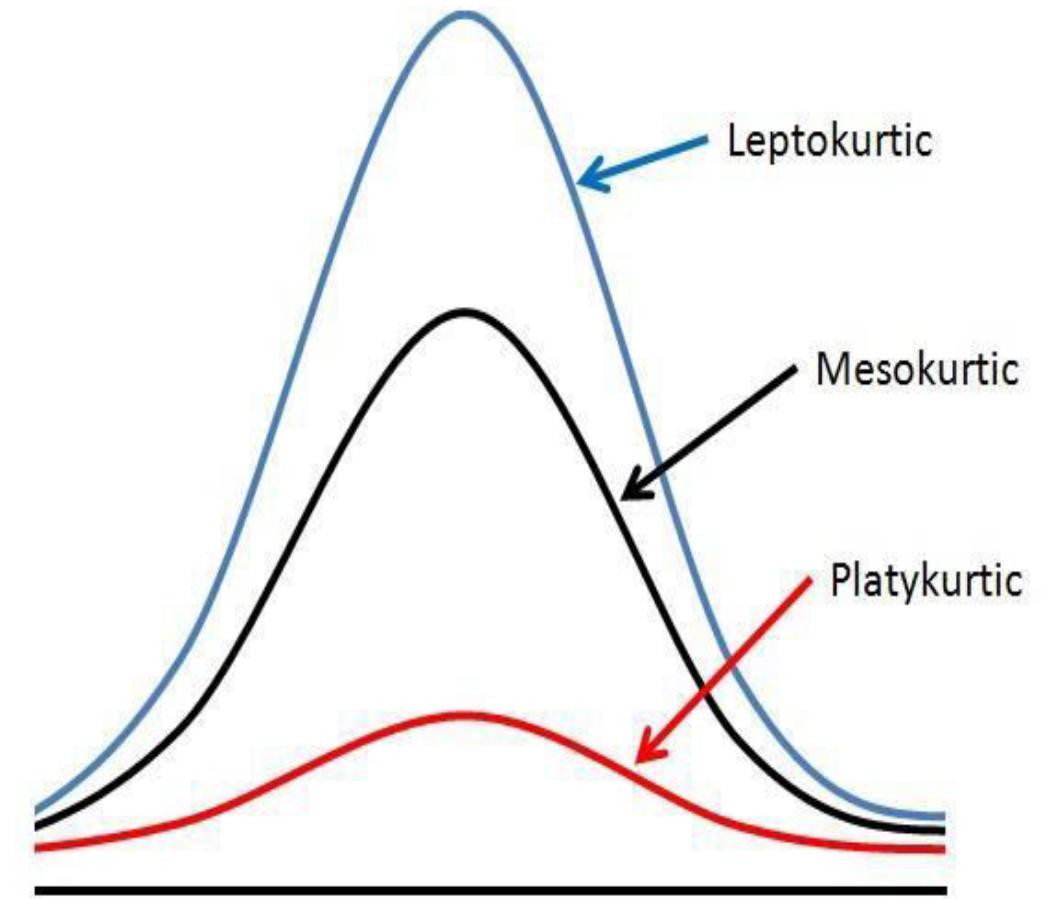
High kurtosis in a data set is an indicator that data have heavy tails or outliers. Low kurtosis in a data set is an indicator that data has light tails or lack of outliers.

Leptokurtic, mesokurtic and platykurtic distributions

Mesokurtic: This distribution has kurtosis statistic similar to that of the normal distribution. It means that the extreme values of the distribution are similar to that of a normal distribution characteristic. This definition is used regardless of the shape of the distribution.

Leptokurtic (Kurtosis > 3): Distribution is longer, tails are fatter. Peak is higher and sharper than Mesokurtic, which means that data are heavy-tailed or profusion of outliers. Outliers stretch the horizontal axis of the histogram graph, which makes the bulk of the data appear in a narrow ("skinny") vertical range, thereby giving the "skinniness" of a leptokurtic distribution.

Platykurtic: (Kurtosis < 3): Distribution is shorter, tails are thinner than the normal distribution. The peak is lower and broader than Mesokurtic, which means that data are light-tailed or lack of outliers.



Excess Kurtosis

- The excess kurtosis is a measure that captures deviation from kurtosis of a normal distribution.

- Excess Kurtosis =
$$\frac{\sum_{i=1}^n (X_i - \bar{X})^4 / n}{\sigma^4} - 3$$
- Excess kurtosis is a useful metric used in the field of pathology.
- With excess kurtosis, any event in question is prone to extreme outcomes.
- A normal distribution has an excess kurtosis of 0.

Test your understanding!

- The value of $\sum_{i=1}^n (X_i - \bar{X})$ is
 - a) Zero for any sample
 - b) Zero for population but not necessarily for samples
 - c) Zero for both samples and population
 - d) Cannot say

Zero for both samples and population

- Kurtosis indicates how much data resides in the _____ of the distribution
Tail

Test your understanding!

- Mean is greater than median in what kind of distribution?

Positively Skewed Distribution

- In a dataset with 60 observations, 3 parameters were estimated. What is the degrees of freedom?

57

- Calculate Q3 for the data [10,50,30,20,10,20,70,30]

45 $Q_3 = \left(\frac{3(n + 1)}{4} \right)^{\text{th}}$ value of the observation

Solution: $= 6^{\text{th}}$ observation + 0.75 $\left[7^{\text{th}} - 6^{\text{th}} \right]$

$$=(6.75)^{\text{th}} \text{ observation}$$

$$=30+0.75(20)=45$$

References

- Business Analytics by U. Dinesh Kumar – Wiley 2nd Edition, 2022
Chapter : 2.6 – 2.8



—
S
TY

Solve

1. The daily footfall at a retail store in Bangalore over the last 30 days is shown in Table 2.7. Calculate the mean, median, mode and standard deviation.

TABLE 2.7 Footfall data

232	277	261	173	283	197	251	212	213	213
229	164	219	196	186	247	244	269	216	272
252	314	161	165	221	260	219	290	225	251

2. For the data in Table 2.7, calculate the skewness and kurtosis. What can you infer from the skewness and kurtosis of the footfall data?
3. For the data in Table 2.7, calculate the values of first quartile and third quartile. Are there any outliers in the data?

TABLE 2.8 CGPA of students

3.36	1.56	1.48	1.43	2.64	1.48	2.77	2.20	1.38	2.84
1.88	1.83	1.87	1.95	3.43	1.28	3.67	2.23	1.71	1.68
2.57	3.74	1.98	1.66	1.66	2.96	1.77	1.62	2.74	3.35
1.80	2.86	3.28	1.14	1.98	2.96	3.75	1.89	2.16	2.07

- (a) Calculate the mean, median and mode. Calculate the standard deviation.
- (b) Calculate the 90th and 95th percentile of CGPA.
- (c) Calculate the inter quartile range (IQR).
- (d) The Dean of the school believes that the CGPA is a right tailed distribution. Is there an evidence to support dean's belief?



THANK YOU

Dr. Gowri Srinivasa

Professor, Department of Computer Science
and Engineering, PES University, Bengaluru

Email: gsrinivasa@pes.edu





PES
UNIVERSITY

DATA ANALYTICS

UE21CS342AA2

UNIT-1

Lecture 4 : Data Preprocessing - Cleaning

Gowri Srinivasa

Department of Computer Science and Engineering

Data Analytics

Unit 1

Lecture 4 : Data Preprocessing - Cleaning

Slides excerpted from: Data Mining : Concepts and Techniques by Han, Kamber and Pei, 3rd Edition

Gowri Srinivasa

Department of Computer Science and Engineering

Slides collated by:

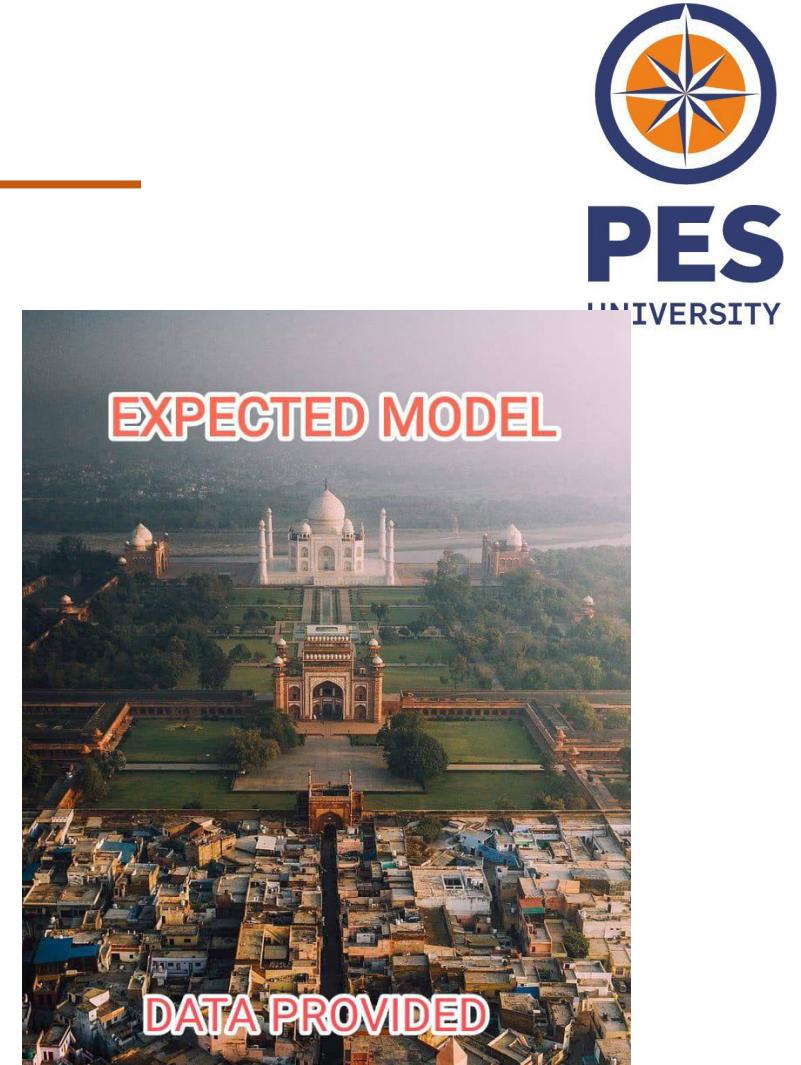
Nishanth M S, VII Sem, Department of CSE, PESU
nishanthmsathish.23@gmail.com

Harshitha Srikanth, VII CSE, PES University
harshithasrikanth13@gmail.com

With grateful thanks for contribution of slides to:
Dr. Mamatha H R, Professor at the Department of CSE, PESU

Data Preprocessing

- Analysis on data can only be as good as the data itself. Low quality data will lead to low quality analysis.
- Real world databases are highly susceptible to noisy , missing and inconsistent data owing to their huge size and multiple heterogeneous sources.
- Data processing techniques when applied before analysis can substantially improve the overall quality of analysis and/or the time required for the actual analysis.



Measures of Data Quality

1) **Accuracy** : Data must not contain errors or a lot of noise.

Example of inaccurate data : Date = 30/02/2002.

Reasons for inaccurate data :

- Data collection instruments may be faulty.
- Human errors occur during data entry.
- Disguised missing data : Users may purposefully submit incorrect data values for mandatory fields when they don't want to share their personal information. Example : Choosing the default value of January 1st for date of birth.

Measures of Data Quality

2) **Completeness** : Data must not lack attribute values. It must contain attributes of interest and relevance to the problem at hand.

Reasons for incompleteness :

- Attributes of interest were not considered important at the time of entry.
- Data might not be recorded due to equipment malfunction resulting in missing data.

3) **Consistency** : Should not contain any discrepancies in the data or the naming convention of the attributes.

Examples of inconsistency:

- Age is recorded as 50 but Date of Birth = 03/04/2005.
- In the result column of students' marks , few entries are in GPA format and rest in percentage.
- Discrepancies can exist between duplicate records.

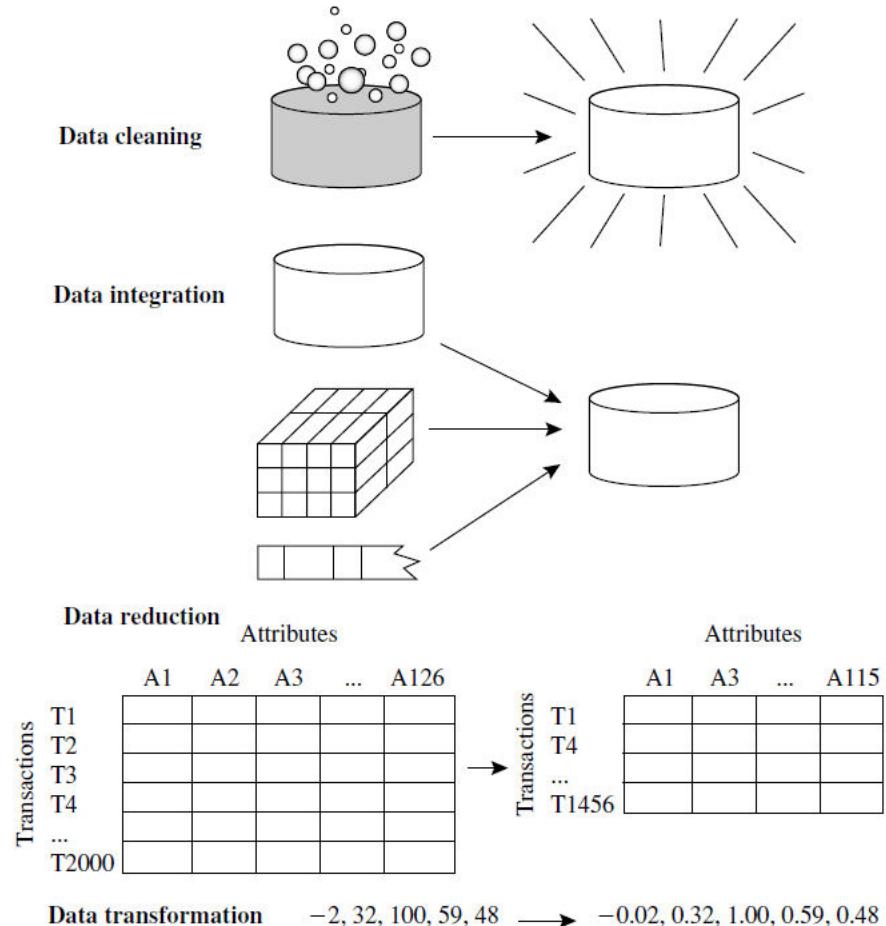
Measures of Data Quality

- 4) **Timeliness** : The data must be updated in a timely fashion. For example , for an analysis run on the first day of every month, previous month's data must be up to date for accurate analysis.
- 4) **Interpretability** : The data must be easily understood. If the attributes of the data aren't easily understandable , the analysis is going to be hindered.
- 4) **Believability** : The data and its source must be trusted by the users. If this data or the source caused problems in the past , current users will find it hard to trust it.

NOTE : The quality of data is subjective and depends on the intended use of data. The data needs of each problem is different.

Major Tasks in Data Preprocessing

- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation



Data Cleaning

Data cleaning entails :

- Filling in missing values
- Smoothening noisy data
- Identifying and removing outliers

If users believe the data is dirty, they are unlikely to trust the outcome of the analysis.



Missing Data

Data is not always complete. Missing data maybe due to :

- Equipment malfunction
- Inconsistent with other recorded data leading to its deletion
- Data not recorded due to a misunderstanding
- Certain data may not be considered useful at the time of entry

Missing data may need to be inferred.

Handling missing data

1. **Ignore the tuple** : Usually done when the class label is missing (for a classification task). This is not effective when the percentage of missing values per attribute varies considerably
2. **Fill in the missing value manually** : Time consuming and infeasible for a large data set.
3. **Fill it with a global constant** : Replace it with a global constant like the word “Unknown”. The downside is that the model might learn patterns with respect to the occurrence of the word “Unknown”.
4. **Fill it with a central tendency** : For symmetric data distributions , replace it with the mean and for skewed data distributions, replace it with the median.
5. A smarter way is to use attribute mean or median(based on the distribution)for all samples belonging to the same class.
6. **Use most probable model** : Use models like regression , decision tree or inference-based Bayesian formalism to infer the missing value.

Types of Missing Values

1. Missing Completely At Random (MCAR)

- The missing data is independent of the observed and unobserved data. In other words, no systematic differences exist between records with missing data and those with complete data.
- For example : A weighing scale running out of batteries. This is not dependent on the person and the probability of this happening is equal to everyone.
- Assuming the data as MCAR is a strong and often unrealistic assumption as “true randomness” is rare in the real world.
- MCAR data doesn’t add bias to the analysis.
- Ways to deal with it :
 - Delete the records : If it is a small fraction of data
 - Delete the attributes : If it is a small fraction of attributes
 - Mean imputation
 - Pairwise deletion : Compute the mean, variance and covariance with another variable available.

Types of Missing Values

Missing Completely At Random (MCAR)

Situation: The school handed out paper surveys to 500 students. However, due to a random event, let's say a gust of wind, 50 of these surveys flew out of an open window and were lost.

Why it's MCAR: The 50 surveys that are missing are not related to the students' satisfaction levels or any other characteristic of the students. It was just a random event that caused these specific surveys to go missing. There's no reason to believe, for example, that the lost surveys were more likely to come from students who were particularly satisfied or dissatisfied.

Implications: Since the missing data is MCAR, the 450 surveys that the school has are still a good representation of the entire student body's opinions. Any analysis done on the 450 surveys will likely produce results that are applicable to the full population of 500 students, because the missingness was not related to any systematic bias.

Types of Missing Values

2. Missing At Random (MAR)

- MAR assumes that the missing value can be predicted based **on the other observed data. The missingness is still random.**
- Example : Employed people are less likely to answer all questions of a survey when compared to unemployed people. **Data is MAR if the likelihood of completing the survey is dependent of the employment status but not on the topic of survey.**
- Almost always produces a bias in the analysis.
- MCAR implies MAR but the converse isn't true.
- Ways to deal with it :
 - Regression imputation : unbiased if it considers the factor which influences the missingness.
 - Last observation carried forward (LOCF) and Baseline observation carried forward(BOCF) : Yields biased estimates. Must be used only if the underlying assumptions are scientifically justifiable.
 - Use of multiple imputation (Packages mice and amelia in R)

Types of Missing Values-Handling MAR

- **Regression Imputation:** In this approach, you might use a variable like 'gender' to predict and impute the missing 'income' values. A regression model is trained using the non-missing values and then used to predict the missing values based on the observed data.
- **Last Observation Carried Forward (LOCF)-** LOCF imputes missing data by taking the last observed data point and carrying it forward. For instance, if you have monthly measurements and a participant's data is missing for April and May, but March's data is available, you'd use the March data for both April and May. **Potential Bias:** LOCF assumes that a participant's condition remains unchanged after their last observation. This can introduce bias if, for example, the true trend is for condition improvement or deterioration.
- **Baseline Observation Carried Forward (BOCF)-** BOCF imputes missing data by taking the baseline (usually the first observed data point) and using it for any subsequent missing points. **Potential Bias:** BOCF assumes that a participant returns to their baseline condition in the event of a missing observation. This can be problematic especially if the treatment or intervention has any enduring effect.

Types of Missing Values

2. Missing At Random (MAR)

Scenario: School Survey on Student Well-being

The school distributes a comprehensive survey to students to gather data on their well-being, including questions on mental health, physical health, academic stress, and participation in extracurricular activities.

Observation:

- You notice that many students skipped questions related to mental health, but you also observe that students who reported not participating in any extracurricular activities are more likely to have skipped the mental health questions.
- "missingness isn't dependent on the actual (unobserved) mental health status of the students" in the MAR example means that whether or not a student chooses to answer the mental health questions isn't directly related to their true mental health status. Instead, the decision to skip the questions is related to another observed variable, in this case, participation in extracurricular activities.

Types of Missing Values

3. Missing Not At random (MNAR)

- The missingness of the data depends on the value of the data. The mechanism for why the data is missing is known. Yet , the values can't be effectively inferred.
- Examples :
 - Censored data
 - People belonging to certain income brackets might not wish to disclose their assets.
 - A weighing machine can only measure weights in a particular range.
- Ways to deal with this :
 - One must model the missingness explicitly, jointly modelling the response and missingness.
 - Generally , the data is assumed to be MAR whenever feasible to avoid this situation.

NOTE : There is no statistical way to determine under which category your missing data will fall under.

Types of Missing Values

Missing Not At random (MNAR)

If the students who were experiencing mental health challenges were specifically the ones avoiding the mental health questions, it would be an example of Missing Not At Random (MNAR). This is because the missingness would be directly dependent on the unobserved, true value of the missing data itself.

Types of Missing Values-A Quick Glance

Missing Completely at Random, MCAR, means there is no relationship between the missingness of the data and any values, observed or missing. Those missing data points are a random subset of the data. There is nothing systematic going on that makes some data more likely to be missing than others.

Missing at Random, MAR, means there is a systematic relationship between the propensity of missing values and the observed data, but not the missing data.

Whether an observation is missing has nothing to do with the missing values, but it does have to do with the values of an individual's observed variables. So, for example, if men are more likely to tell you their weight than women, weight is MAR.

Missing Not at Random, MNAR, means there is a relationship between the propensity of a value to be missing and its values.

An Interesting Thought



- Imagine you are collecting some information from your classmates. For many reasons , not everyone will answer every question of yours. And that is okay!
- Well the next step is replacing missing values right? We can use any one of the methods we have discussed till now after some analysis of the data.
- But wait! Don't you think the fact that they did not answer is some kind of information per se which can be beneficial to our analysis?
- So the next time you build a model , before dealing with the missing values , create an additional variable (preferably a binary variable) in which you store if the particular student answered or not.
- This may (or may not!) help you gain more insights about the population or improve the analytics model you are building!

Noisy data

Noise is a random error or variance in a measured variable.

Data smoothening techniques to combat noise :

- **Binning**
 - Sort the data and partition into bins(equal-width, equal-frequency, etc.)
 - Smooth by bin means, bin medians, by bin boundaries etc.
 - More on binning in further lectures.
- **Regression** - Data can be smoothed by fitting it to a regression model.
- **Clustering** - Outliers can be detected with the help of clustering and can be removed to smoothen the data.
- **Combined computer and human inspection** – Computer detects suspicious values and is validated by a human. Is useful when dealing with possible outliers.

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

Partition into (equal-frequency) bins:

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

Smoothing by bin means:

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

Smoothing by bin boundaries:

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

Figure 3.2 Binning methods for data smoothing.

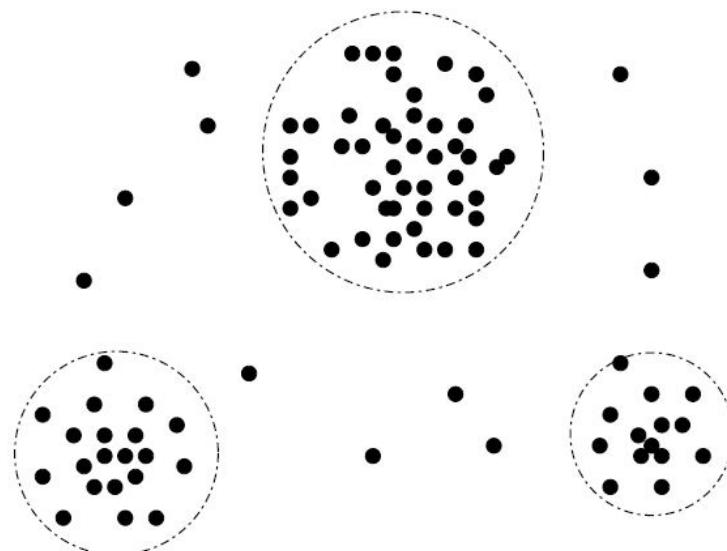
Outliers

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 main goal of our analysis. Examples :

- Credit card fraud
- Spam detection
- Intrusion detection



Data Cleaning as a Process

- Data discrepancy detection
 - Refer to the metadata of data to gain knowledge regarding its properties.
 - Perform summary statistics for all attributes and discover the distributions, dependencies , outliers and so on.
 - Look for inconsistent representation of data. For example , make sure all dates are following the same format , for instance , DD-MM-YYYY.
 - Check for field overloading – practice of coupling two or more data elements to a single field. It ensures efficient memory utilization.
 - Check for uniqueness rule, consecutive rule and null rule.
 - Uniqueness rule : Each value of the given attribute must be unique.
 - Consecutive rule : There can't be any missing values between lowest and highest value for that attribute. All values must be unique. Example - cheque number.
 - Null rule : Specifies how to record a null value. For example , use 0 for numeric attribute and ‘?’ for nominal attribute.

Data Cleaning as a Process

- Data discrepancy detection
 - Use commercial tools that can aid in this step.
 - Data scrubbing tools : Use simple domain knowledge (example, knowledge of postal zip code and spell-check) to detect errors and make corrections.
 - Data auditing tools : Find discrepancies by analyzing the data to discover rules and relationships , and detect data that violates the discovered rules. For example , it employs statistical analysis to find correlations or clustering to detect outliers.

Data Cleaning as a Process

- Data transformation
 - Some data inconsistencies can be corrected manually but most errors require data transformations.
 - Data migration tools allow transformations to be specified.
 - ETL (Extraction/Transformation>Loading) tools allow users to specify transformations through a graphical user interface.
- Data transformations may introduce more discrepancies.
- The 2-step process of discrepancy detection and data transformation occurs iteratively until no further anomalies are found.
- New approaches to data cleaning emphasize increased interactivity. Potter's wheel is a publicly available data cleaning tool that integrates both the steps.

Test your understanding!

- Which of these is **not** a method to deal with noisy data?
 - a) Binning
 - b) Regression
 - c) Principal Component Analysis
 - d) Clustering

Solution

c) Principle Component Analysis

- Outliers need to be removed in every dataset , regardless of the problem statement.

Solution

False

- Mean imputation can be done for which type of missing data?

Solution

MCAR

Test your understanding!

- The statement “Most of the missing people from work are sickest people” denotes what type of missingness?

MNAR

- Which type of missingness is called “non-ignorable”?

MNAR

Because the missing data mechanism itself has to be modelled as you deal with the missing data. You have to include some model for why the data are missing and what the likely values are.

References

- [Data Mining : Concepts and Techniques](#) by Han, Kamber and Pei , The Morgan Kaufmann Series in Data Management Systems ,3rd Edition Chapter : 3.1-3.2
- <http://dept.stat.lsa.umich.edu/~jerrick/courses/stat701/notes/mi.html>
- <https://www.scribbr.com/statistics/missing-data/>
- <https://www.theanalysisfactor.com/missing-data-mechanism/>



THANK YOU

Dr. Gowri Srinivasa

Professor, Department of Computer Science
and Engineering, PES University, Bengaluru

Email: gsrinivasa@pes.edu



DATA ANALYTICS

UE20CS312

UNIT-1

Lecture 5 : Data Preprocessing
- Data Integration and Reduction

Gowri Srinivasa

Department of Computer Science and Engineering

Data Analytics

Unit 1

Lecture 6 : Data Preprocessing – Data Integration and Reduction

Slides excerpted from: Data Mining : Concepts and Techniques by Han, Kamber and Pei, 3rd Edition

Gowri Srinivasa

Department of Computer Science and Engineering

Slides collated by:

Nishanth M S PESU-2023, Department of CSE

nishanthmsathish.23@gmail.com

Harshitha Srikanth ,VII Sem ,PESU,Department of CSE

harshithasrikanth13@gmail.com

With grateful thanks for contribution of slides to:

Dr. Mamatha H R, Professor at the Department of CSE, PESU

Data Integration

- Data analysis often requires data integration – the merging of data from multiple data stores into a coherent store.
- Careful integration can help reduce and avoid redundancies and inconsistencies in the resulting dataset. This can help improve the accuracy and speed of the subsequent data analysis process.
- The semantic heterogeneity and structure of data pose great challenges in data integration.
- How can we match schema and objects from different sources?
- **Schema Integration!**
 - Example : How can a data analyst be sure that the attribute customer_id in table A and customer_number in table B refer to the same attribute?
 - With the help of metadata! It provides all possible information regarding the attributes , thus ensuring error free schema integration.

Data Integration

- **Entity identification problem :** Identify real world entities from multiple data sources. Example : Bill Clinton = William Clinton
- **Detecting and resolving data value conflicts**
 - For the same real world entity, attribute values from different sources are different.
 - Possible reasons : different representations, different scales, example – metric vs British units
- During integration , special attention must be paid to the structure of the data. This is to ensure that any attribute functional dependencies and referential constraints in the source system match those in the target system. For example , in one system, a discount may be applied to the entire order whereas in another system , it is applied to each individual line item. If this is not caught before integration, items in the target system may be improperly discounted.

Redundancy in Data Integration

Redundant data often occur during the integration of multiple databases.

- Object identification : The same attribute or object may have different names in different databases which causes redundancy.
- Derivable data : An attribute may be redundant if it can be *derived* from another attribute or set of attributes. For example , annual revenue can be derived from monthly revenue.
- **Few redundancies can be detected by correlation analysis.**
- **For nominal data , χ^2 (chi-square) test is employed.**
- **For numeric data , correlation coefficient and covariance is used.**

χ^2 (chi-square) test

χ^2 (chi-square) test for independence of two variables in a contingency table

- Null Hypothesis : The two variables are independent
- Alternate hypothesis : The two variables are not independent.

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

- *Expected* stands for what we would *expect* if the null hypothesis were true.
- Larger the value of χ^2 the more likely the variables are correlated.
- The cells that contribute the most to the χ^2 value are those whose actual count is different from the expected count.
- Can be used for categorical variables where entries are numbers(counts) and not percentages or fractions(10% of 100 needs to be entered as 10)
- Correlation does not imply causation.
 - The number of hospitals and number of car-thefts in a city may *appear* to be correlated. Both are casually linked to a third variable : population.

DATA ANALYTICS

χ^2 (chi-square) Example

	Play chess	Not play chess	Sum (row)		Play chess	Not play chess	Sum (row)
Like science fiction	250	200	450	Like science fiction	90	360	450
Not like science fiction	50	1000	1050	Not like science fiction	210	840	1050
Sum(col.)	300	1200	1500	Sum(col.)	300	1200	1500
Actual distribution (observed)		Expected distribution					

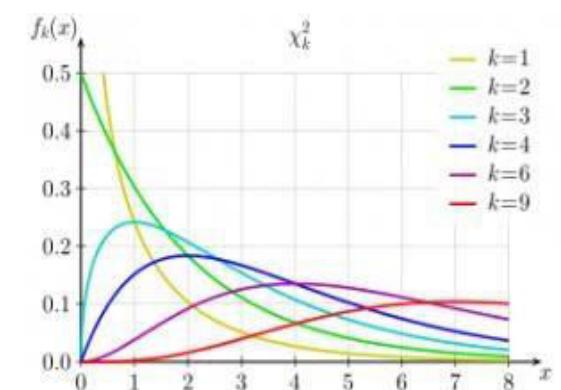
- χ^2 (chi-square) calculation

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

- Degrees of freedom , $k = (\text{no_of_rows}-1)(\text{no_of_columns}-1)=1$
- It shows that like_science_fiction and play_chess are correlated.

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

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



Correlation Analysis (Numeric Data)

- Correlation coefficient (also called as Pearson's product moment coefficient)

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

where n is the number of tuples, \bar{A} and \bar{B} are the respective means of A and B , σ_A and σ_B are the respective standard deviations of A and B , and $\Sigma(a_i b_i)$ is the sum of the AB cross-product.

- If $r_{A,B} > 0$, A and B are positively correlated , that is A's values increase when B's does. The higher the value of coefficient , stronger the correlation.
- If $r_{A,B} = 0$: A and B are independent of each other.
- If $r_{A,B} < 0$: A and B are negatively correlated. That is A's values decrease when B's increases.

Correlation (viewed as a linear relationship)

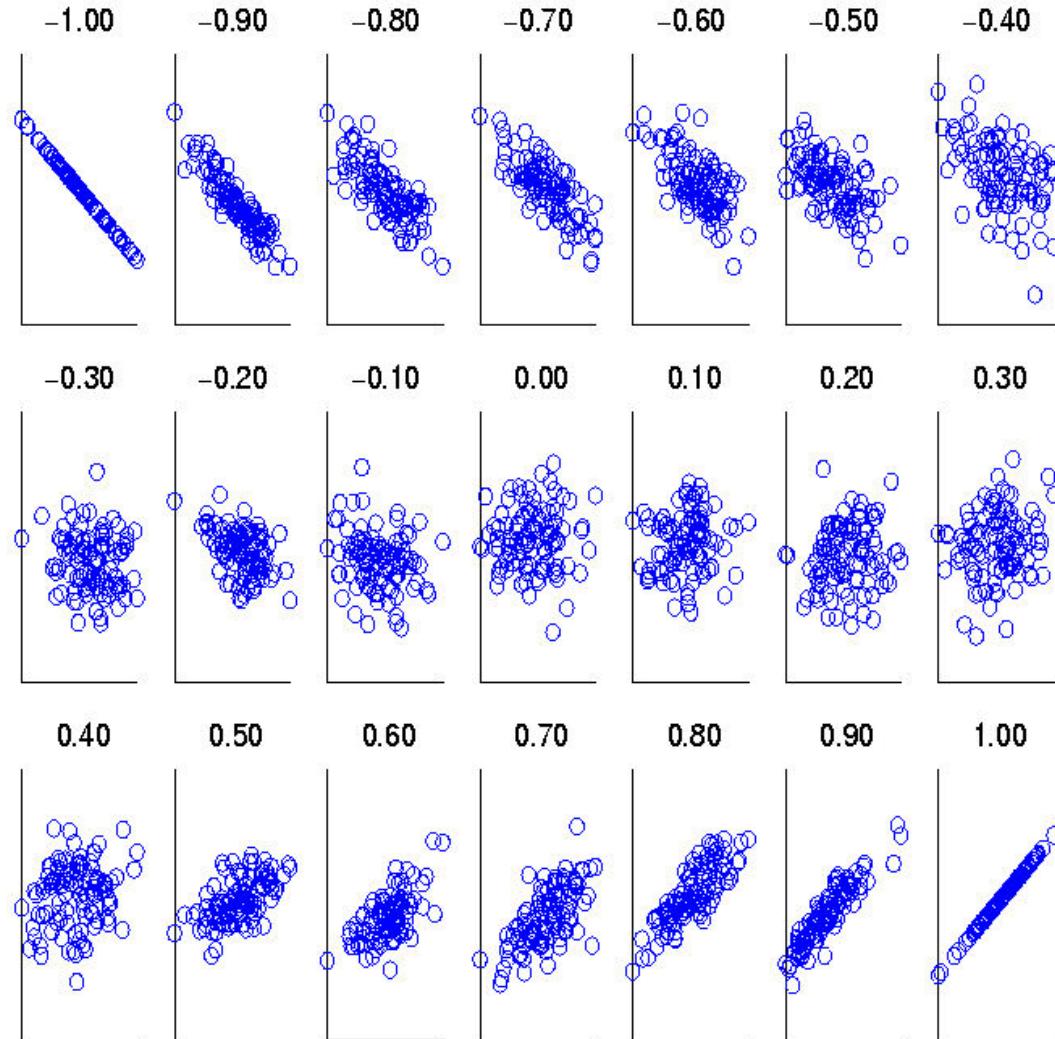
- Correlation measures the linear relationship between objects.
- To compute correlation , we **standardize data objects A and B (vector)**, and then take their dot product.

$$a'_k = (a_k - \text{mean}(A)) / \text{std}(A)$$

$$b'_k = (b_k - \text{mean}(B)) / \text{std}(B)$$

$$\text{correlation}(A, B) = A^T B$$

Visually Evaluating Correlation



Scatter plots signifying
the strength of
correlation

Covariance analysis (Numeric Data)

- Covariance is similar to correlation

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

Correlation coefficient:

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

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

Covariance analysis (Numeric Data)

- **Positive Covariance :** If $\text{Cov}_{A,B} > 0$, then A and B both tend to be larger than their expected values. (A **positive covariance** indicates that as one variable increases, the other tends to increase as well.)
- **Negative Covariance :** If $\text{Cov}_{A,B} < 0$, then if A is larger than its expected value, B is likely to be smaller than its expected value.(A negative covariance indicates that as one variable increases, the other tends to decrease.)
- **Independence :** $\text{Cov}_{A,B} = 0$, but the converse is not true :A covariance of zero suggests that the variables are not linearly related

Some pairs of random variables may have a covariance of 0 but are not independent. Only under few additional assumptions (example , the data follows multivariate normal distributions) does $\text{Cov}_{A,B} = 0$ imply independence.

Independence:

Two random variables A and B are said to be independent if knowing the outcome of one doesn't give any information about the outcome of the other. Zero Covariance ≠ Independence

Covariance analysis : An Example

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

- It can be simplified in computation as

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

Covariance analysis : An Example

Suppose two stocks A and B have the following values in one week : (2, 5), (3, 8), (5, 10), (4, 11), (6, 14).

Question: If the stocks are affected by the same industry trends, will their prices rise or fall together?

$$E(A) = (2 + 3 + 5 + 4 + 6) / 5 = 20/5 = 4$$

$$E(B) = (5 + 8 + 10 + 11 + 14) / 5 = 48/5 = 9.6$$

$$\text{Cov}(A,B) = (2 \times 5 + 3 \times 8 + 5 \times 10 + 4 \times 11 + 6 \times 14) / 5 - 4 \times 9.6 = 4$$

Thus, A and B rise together since $\text{Cov}(A, B) > 0$.

Tuple Duplication

- In addition to detecting redundancies between attributes, duplication should be detected at the tuple level (Example , where there are two or more identical tuples for a unique data entry case)
- The use of denormalized tables (often done to improve performance by avoiding joins) is another source of data redundancy.
- Inconsistencies often arise between various duplicates, due to inaccurate data entry or updating some but not all data occurrences.

Data Value Conflict Detection and Resolution

- Data integration also involves the detection and resolution of data value conflicts.
- For example, for the same real-world entity, attribute values from different sources may differ.
- This may be due to differences in representation, scaling or encoding.
- For instance , a weight attribute may be stored in metric units in once system and British imperial units in another.

Data Reduction

- Data reduction techniques are applied to obtain a reduced representation of the dataset that is much smaller in volume , yet closely maintains the integrity of the original data.
- Analysis on the reduced dataset should be more efficient yet produce the same or almost the same analytical results.
- Why do we need data reduction? A database or a data warehouse may store terabytes of data. Complex data analysis may take a very long time to run on the complete data set.

Data Reduction Strategies

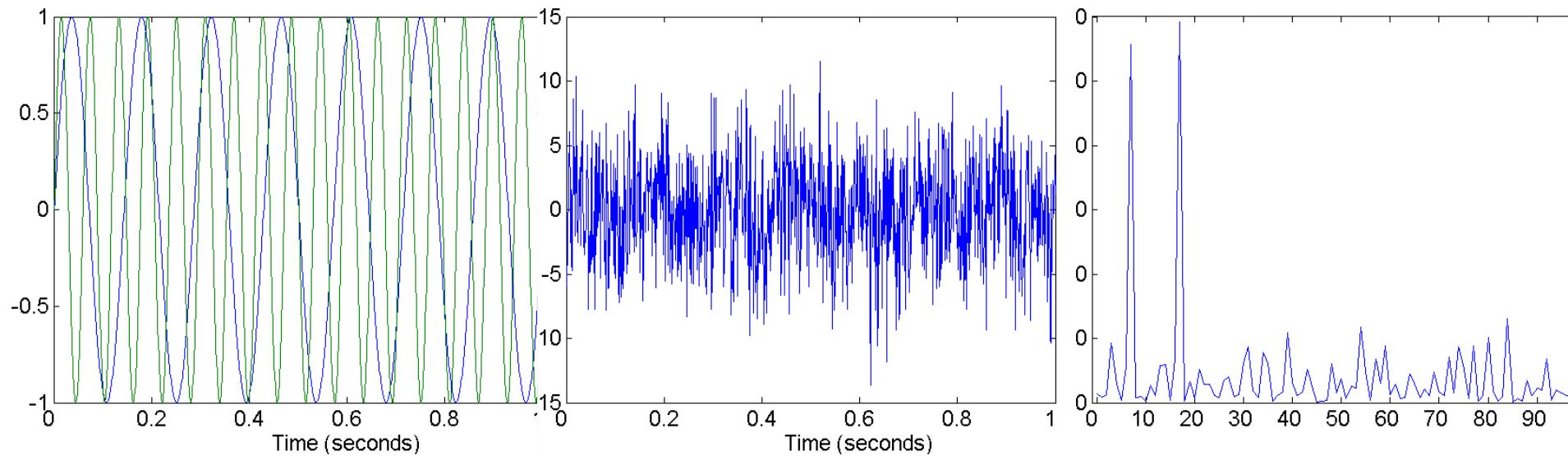
- Dimensionality reduction – process of removing unimportant attributes
 - Wavelet transforms
 - Principal Component Analysis (PCA)
 - Attribute subset selection
- Numerosity reduction – replaces the original data volume by an alternative, smaller forms of data representation
 - Regression and log-linear models
 - Histograms, clustering and sampling
 - Data cube aggregation
- Data compression – transformations are applied on the data to obtain a *reduced* or a *compressed* representation of the original data.

Dimensionality Reduction

- Curse of dimensionality
 - When dimensionality increases, data becomes increasingly sparse.
 - Density and distance between points , which are critical to clustering and outlier analysis become less meaningful.
 - The possible combinations of subspaces will grow exponentially.
- Dimensionality reduction
 - Avoids the curse of dimensionality.
 - Helps to eliminate irrelevant attributes and reduce noise.
 - Reduces time and space required for data analytics.
 - Enables easier visualization.

Mapping data to a new space

- Fourier transform
- Wavelet transform



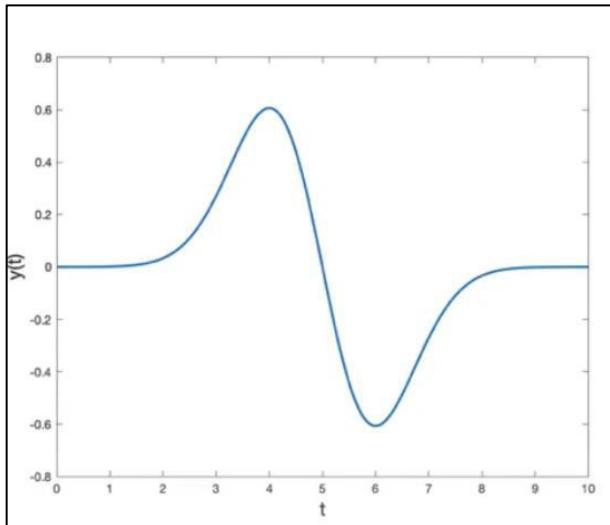
Two Sine Waves

Two Sine Waves + Noise

Frequency

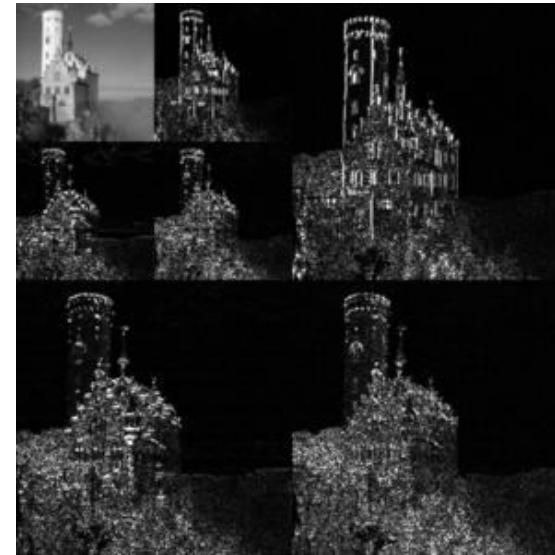
What is a Wavelet?

A **Wavelet** is a **wave-like oscillation that is localized in time**, an example is given below. Wavelets have two basic properties: scale and location. **Scale** (or dilation) defines how “stretched” or “squished” a wavelet is. This property is related to frequency as defined for waves. **Location** defines where the wavelet is positioned in time (or space).

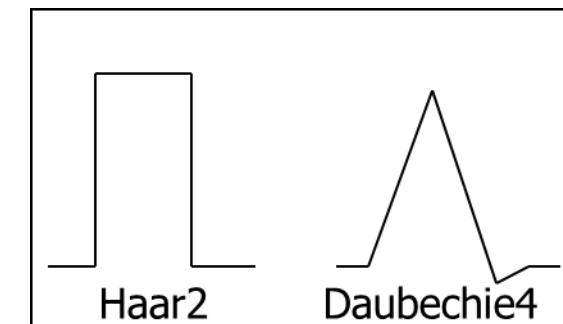


Wavelet transformation

- Discrete wavelet transform (DWT) is used for linear signal processing and multi-resolution analysis.
- It decomposes a signal into different frequency sub-bands. It is applicable to n-dimensional signals.
- Data is transformed to preserve relative distance between objects at different resolutions.
- Compressed approximation : it stores only a small fraction of the strongest of the wavelet coefficients
- It is insensitive to noise , input order and is only applicable to low dimensional data.



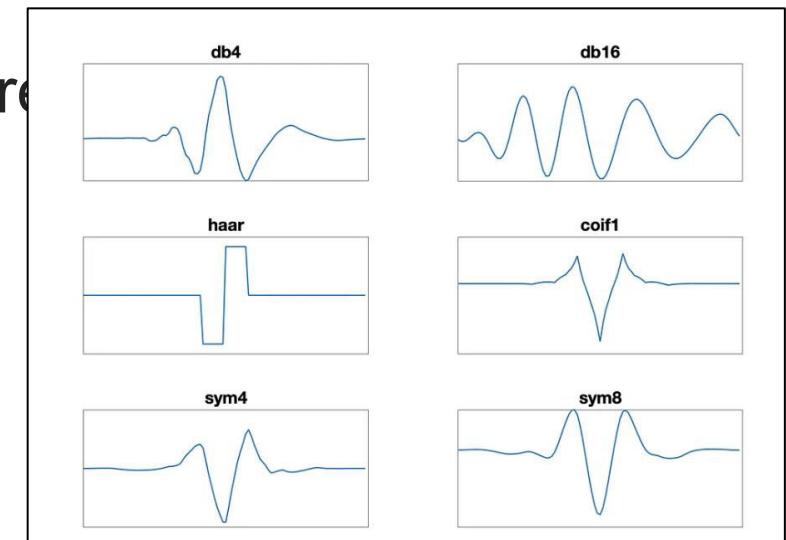
An example of DWT



Wavelet families

Why wavelet transforms?

- A major disadvantage of the Fourier Transform is it captures *global* frequency information, meaning frequencies that persist over an entire signal. An alternative approach is the **Wavelet Transform**, which **decomposes a function into a set of wavelets**.
- **Wavelet transform** can extract local spectral **and** temporal information simultaneously
- **Variety of wavelets** to choose from like shown here

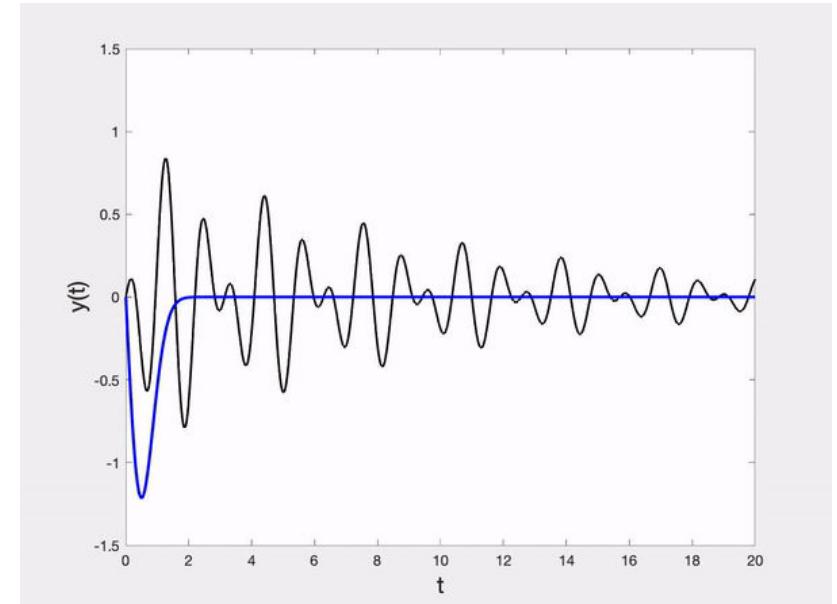


Wavelet transformation-Working

Take a look at the animation first.

Now, the basic idea is to compute "how much" of a wavelet is in a signal for a particular scale and location. For those familiar with convolutions, that is exactly what this is. A signal is convolved with a set wavelets at a variety of scales.

In other words, we pick a wavelet of a particular scale (like the blue wavelet in the gif). Then, we slide this wavelet across the entire signal i.e. vary its location, where at each time step we multiply the wavelet and signal. The product of this multiplication gives us a coefficient for that wavelet scale at that time step. We then increase the wavelet scale (e.g. the red and green wavelets) and repeat the process.



Wavelet transformation-Working

Like the Fourier transform, the wavelet transform deconstructs the original signal waveform into a series of basis waveforms, which in this case are called *wavelets*. However, unlike the simple sinusoidal waves of Fourier analysis, the wavelet shapes are complex, and, at first sight apparently arbitrary – they look like random squiggles (although in fact they fulfil rigorous mathematical requirements). One important feature that all wavelets share is that they are *bounded*, i.e. they decline to zero amplitude at some distance either side of the centre, which is in obvious contrast to the sine/cosine waves used in Fourier analysis, which go on forever. This is the underlying key to the time localisation of the DWT.

There are a whole series of different types of “mother” wavelets (Daubechies, Coiflet, Symmlet etc) available, and each type occurs in a range of sizes. A particular episode of wavelet analysis only uses one type of mother wavelet; the user decides which type and size to use depending on the characteristics of the signal to be analysed (and probably some trial-and-error)

Wavelet transformation-Working

After transformation of a raw data signal using a particular mother wavelet you end up with basis waveforms consisting of a series of daughter wavelets. The daughter wavelets are all compressed or expanded versions of the mother wavelet (they have different *scales* or frequencies), and each daughter wavelet extends across a different part of the original signal (they have different *locations*)

The important point is that each daughter wavelet is associated with a corresponding *coefficient* that specifies how much the daughter wavelet at that scale contributes to the raw signal at that location. It is these coefficients that contain the information relating to the original input signal, since the daughter wavelets derived from a particular mother wavelet are completely fixed and independent of the input signal. Like the Fourier transform, the wavelet transform is reversible - you can reconstruct the original signal by adding together the appropriately daughter wavelets, each weighted by its associated coefficient.

Wavelet transformation-Working

“How can this technique be useful for data reduction if the wavelet transformed data are of the same length as the original data?”

The usefulness lies in the fact that the wavelet transformed data can be truncated. A compressed approximation of the data can be retained by storing only a small fraction of the strongest of the wavelet coefficients.

For example, all wavelet coefficients larger than some user-specified threshold can be retained. All other coefficients are set to 0. The resulting data representation is therefore

very sparse, so that operations that can take advantage of data sparsity are computationally very fast if performed in wavelet space. The technique also works to remove noise without smoothing out the main features of the data, making it effective for data cleaning as well.

References

- [Data Mining: Concepts and Techniques](#) by Jiawei Han, Micheline Kamber and Jian Pei, The Morgan Kaufmann Series in Data Management Systems, 3rd Edition Chapter : 3.3 – 3.4
- <https://www.st-andrews.ac.uk/~wjh/dataview/tutorials/dwt.html>
- <https://towardsdatascience.com/the-wavelet-transform-e9cfa85d7b34>
- https://www.cs.unm.edu/~mueen/Teaching/CS_521/Lectures/Lecture2.pdf
- <https://medium.com/analytics-vidhya/understanding-principle-component-analysis-pca-step-by-step-e7a4bb4031d9>



THANK YOU

Dr. Gowri Srinivasa

Professor, Department of Computer Science
and Engineering, PES University, Bengaluru

Email: gsrinivasa@pes.edu



DATA ANALYTICS

UE21CS342AA2

UNIT-1

**Lecture 6 : Data and Dimensionality reduction
contd.**

Gowri Srinivasa

Department of Computer Science and Engineering

Principal Component Analysis (PCA)

What is PCA?

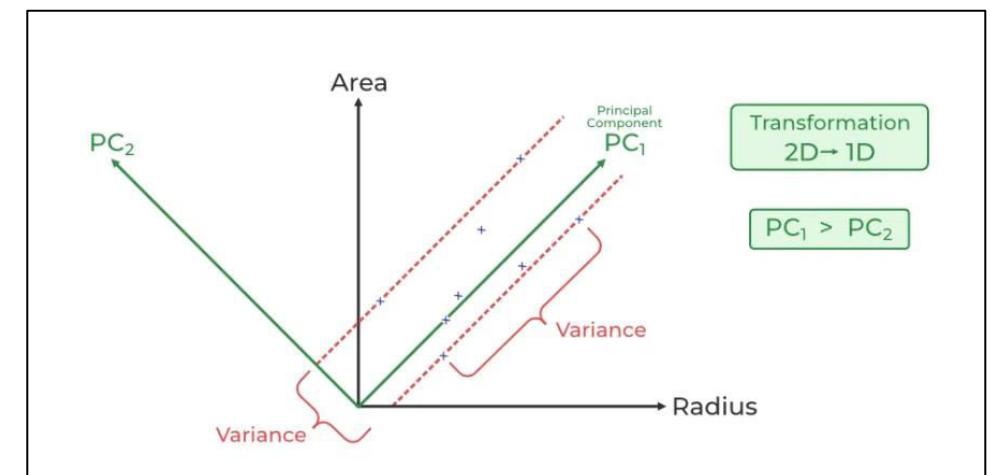
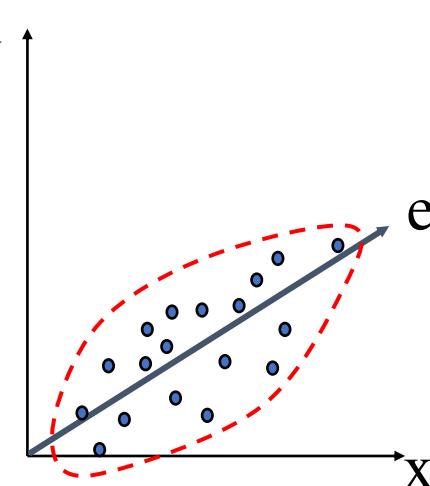
Assume there are 50 questions in all in the survey. The following three are among them:

- 1.I feel comfortable around people
- 2.I easily make friends
- 3.I like going out

These queries could appear different now. There is a catch, though. They aren't, generally speaking. They all gauge how extroverted you are. Therefore, combining them makes it logical, right? That's where linear algebra and dimensionality reduction methods come in! We want to lessen the complexity of the problem by minimizing the number of variables since we have much too many variables that aren't all that different. That is the main idea behind dimensionality reduction. And it just so happens that PCA is one of the most straightforward and popular techniques in this field.

Principal Component Analysis (PCA)-lossy

- Finds a projection that captures the largest amount of variation in data.
- The original data is projected onto much smaller space , resulting in dimensionality reduction.
- We find eigenvectors of the covariance matrix and these eigenvectors define the new space.



PCA - Steps

Step 1: Standardize the dataset.

Step 2: Calculate the covariance matrix for the features in the dataset.

Step 3: Calculate the eigenvalues and eigenvectors for the covariance matrix.

Step 4: Sort eigenvalues and their corresponding eigenvectors.

Step 5: Pick k eigenvalues and form a matrix of eigenvectors.

Step 6: Transform the original matrix

Lets go step by step

PCA - Steps

1. Standardize the Dataset

Assume we have the below dataset which has 4 features and a total of 5 training examples.

f1	f2	f3	f4
1	2	3	4
5	5	6	7
1	4	2	3
5	3	2	1
8	1	2	2

Dataset matrix

First, we need to standardize the dataset and for that, we need to calculate the mean and standard deviation for each feature.

$$x_{new} = \frac{x - \mu}{\sigma}$$

Standardization formula

	f1	f2	f3	f4
$\mu =$	4	3	3	3.4
$\sigma =$	3	1.58114	1.73205	2.30217

Mean and standard deviation before standardization

After applying the formula for each feature in the dataset is transformed as below:

f1	f2	f3	f4
-1	-0.63246	0	0.26062
0.33333	1.26491	1.73205	1.56374
-1	0.63246	-0.57735	-0.17375
0.33333	0	-0.57735	-1.04249
1.33333	-1.26491	-0.57735	-0.60812

Standardized Dataset

PCA - Steps

2. Calculate the covariance matrix for the whole dataset

The formula to calculate the covariance matrix:

For Population

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

For Sample

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

Covariance Formula

the covariance matrix for the given dataset will be calculated as below

	f1	f2	f3	f4
f1	var(f1)	cov(f1,f2)	cov(f1,f3)	cov(f1,f4)
f2	cov(f2,f1)	var(f2)	cov(f2,f3)	cov(f2,f4)
f3	cov(f3,f1)	cov(f3,f2)	var(f3)	cov(f3,f4)
f4	cov(f4,f1)	cov(f4,f2)	cov(f4,f3)	var(f4)

Since we have standardized the dataset, so the mean for each feature is 0 and the standard deviation is 1.

$$\text{var}(f1) = ((-1.0-0)^2 + (0.33-0)^2 + (-1.0-0)^2 + (0.33-0)^2 + (1.33-0)^2)/5$$

$$\text{var } (f1) = 0.8$$

$$\text{cov}(f1,f2) =$$

$$((-1.0-0)*(-0.632456-0) +$$

$$(0.33-0)*(1.264911-0) +$$

$$(-1.0-0)*(0.632456-0) +$$

$$(0.33-0)*(0.000000-0) +$$

$$(1.33-0)*(-1.264911-0))/5$$

$$\text{cov}(f1,f2) = -0.25298$$

In the similar way we can calculate the other covariances and which will result in the below covariance matrix

	f1	f2	f3	f4
f1	0.8	-0.25298	0.03849	-0.14479
f2	-0.25298	0.8	0.51121	0.4945
f3	0.03849	0.51121	0.8	0.75236
f4	-0.14479	0.4945	0.75236	0.8

covariance matrix (population formula)

PCA - Steps

3. Calculate eigenvalues and eigen vectors.

An eigenvector is a nonzero vector that changes at most by a scalar factor when that linear transformation is applied to it. The corresponding eigenvalue is the factor by which the eigenvector is scaled.

Let A be a square matrix (in our case the covariance matrix), v a vector and λ a scalar that satisfies $Av = \lambda v$, then λ is called eigenvalue associated with eigenvector v of A.

Rearranging the above equation,

$$Av - \lambda v = 0 ; (A - \lambda I)v = 0$$

Since we have already know v is a non- zero vector, only way this equation can be equal to zero, if

$$\det(A - \lambda I) = 0$$

	f1	f2	f3	f4
f1	0.8 - λ	-0.25298	0.03849	-0.14479
f2	-0.25298	0.8 - λ	0.51121	0.4945
f3	0.03849	0.51121	0.8 - λ	0.75236
f4	-0.14479	0.4945	0.75236	0.8 - λ

$$A - \lambda I = 0$$

Solving the above equation = 0

$$\lambda = 2.51579324, 1.0652885, 0.39388704, 0.02503121$$

Eigenvectors:

Solving the $(A - \lambda I)v = 0$ equation for v vector with different λ values:

$$\begin{pmatrix} 0.800000 - \lambda & -(0.252982) & 0.038490 & -(0.144791) \\ -(0.252982) & 0.800000 - \lambda & 0.511208 & 0.494498 \\ 0.038490 & 0.511208 & 0.800000 - \lambda & 0.752355 \\ -(0.144791) & 0.494498 & 0.752355 & 0.800000 - \lambda \end{pmatrix} \times \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{pmatrix} = 0$$

For $\lambda = 2.51579324$, solving the above equation using Cramer's rule, the values for v vector are

$$v1 = 0.16195986$$

$$v2 = -0.52404813$$

$$v3 = -0.58589647$$

$$v4 = -0.59654663$$

Going by the same approach, we can calculate the eigen vectors for the other eigen values. We can form a matrix using the eigen vectors.

e1	e2	e3	e4
0.161960	-0.917059	-0.307071	0.196162
-0.524048	0.206922	-0.817319	0.120610
-0.585896	-0.320539	0.188250	-0.720099
-0.596547	-0.115935	0.449733	0.654547

cinvectors(4 * 4 matrix)

And Finally!

4. Sort eigenvalues and their corresponding eigenvectors.

Since eigenvalues are already sorted in this case so no need to sort them again.

5. Pick k eigenvalues and form a matrix of eigenvectors

If we choose the top 2 eigenvectors, the matrix will look like this:

	e1	e2
0.161960	-0.917059	
-0.524048	0.206922	
-0.585896	-0.320539	
-0.596547	-0.115935	

Top 2 eigenvectors(4*2 matrix)

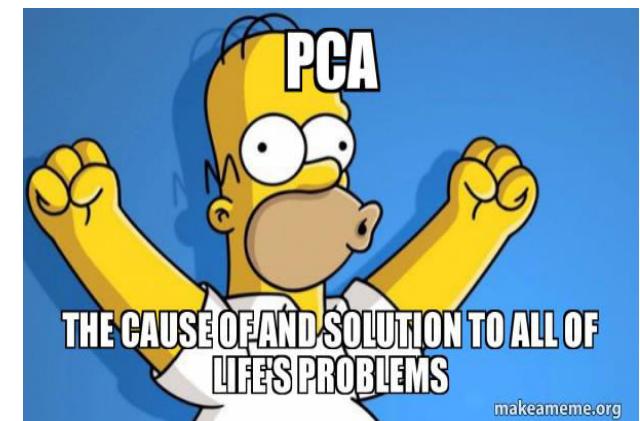
6. Transform the original matrix.

Feature matrix * top k eigenvectors = Transformed Data

$$\begin{array}{cccc} f_1 & f_2 & f_3 & f_4 \\ \hline -1.000000 & -0.632456 & 0.000000 & 0.260623 \\ 0.333333 & 1.264911 & 1.732051 & 1.563740 \\ -1.000000 & 0.632456 & -0.577350 & -0.173749 \\ 0.333333 & 0.000000 & -0.577350 & -1.042493 \\ 1.333333 & -1.264911 & -0.577350 & -0.608121 \end{array} \begin{array}{cc} e_1 & e_2 \\ \hline 0.161960 & -0.917059 \\ -0.524048 & 0.206922 \\ -0.585896 & -0.320539 \\ -0.596547 & -0.115935 \end{array} \begin{array}{cc} nf_1 & nf_2 \\ \hline 0.014003 & 0.755975 \\ -2.556534 & -0.780432 \\ -0.051480 & 1.253135 \\ 1.014150 & 0.000239 \\ 1.579861 & -1.228917 \end{array}$$

(5,4) (4,2) (5,2)

Data Transformation



PCA using R (factoMineR , factoextra)

name	100m	Long.jump	//	Javeline	1500m	Rank	Points	Competition
SEBRLE	11.04	7.58		63.19	291.7	1	8217	Decastar
CLAY	10.76	7.4		60.15	301.5	2	8122	Decastar
Macey	10.89	7.47		58.46	265.42	4	8414	OlympicG
Warners	10.62	7.74		55.39	278.05	5	8343	OlympicG
\\"								
Zsivoczky	10.91	7.14		63.45	269.54	6	8287	OlympicG
Hernu	10.97	7.19		57.76	264.35	7	8237	OlympicG
Pogorelov	10.95	7.31		53.45	287.63	11	8084	OlympicG
Schoenbeck	10.9	7.3		60.89	278.82	12	8077	OlympicG
Barras	11.14	6.99		64.55	267.09	13	8067	OlympicG
KARPOV	11.02	7.3		50.31	300.2	3	8099	Decastar
WARNERS	11.11	7.6		51.77	278.1	6	8030	Decastar
Nool	10.8	7.53		61.33	276.33	8	8235	OlympicG
Drews	10.87	7.38		51.53	274.21	19	7926	OlympicG

Active individuals

Active variables

Supplementary quantitative variables

Supplementary qualitative variable

Supplementary individuals

<http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/112-pca-principal-component-analysis-essentials/>

PCA using R (factoMineR , factoextra)

- Selecting the principal components

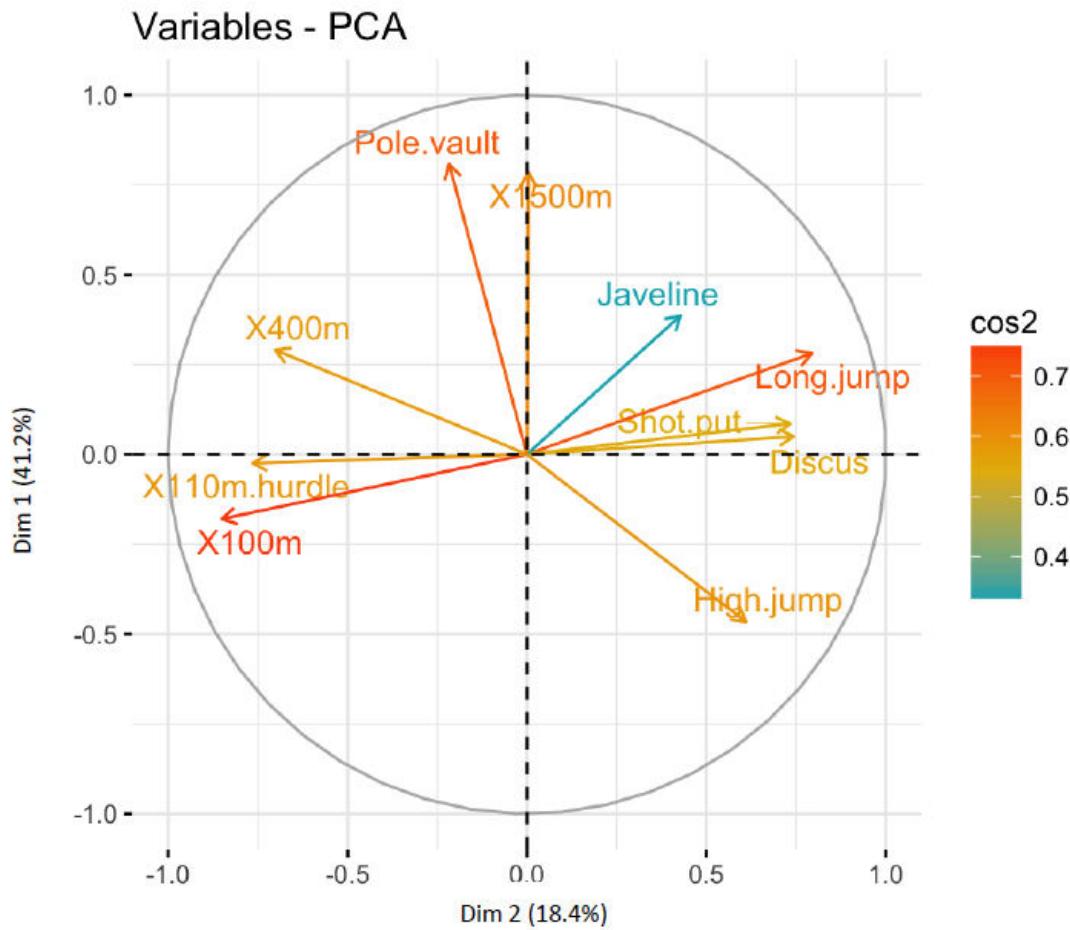
```
library("factoextra") eig.val <-get_eigenvalue(res.pca)  
eig.val
```

	eigenvalue	variance.percent	cumulative.variance.percent
## Dim.1	4.124	41.24	41.2
## Dim.2	1.839	18.39	59.6
## Dim.3	1.239	12.39	72.0
## Dim.4	0.819	8.19	80.2
## Dim.5	0.702	7.02	87.2
## Dim.6	0.423	4.23	91.5
## Dim.7	0.303	3.03	94.5
## Dim.8	0.274	2.74	97.2
## Dim.9	0.155	1.55	98.8
## Dim.10	0.122	1.22	100.0

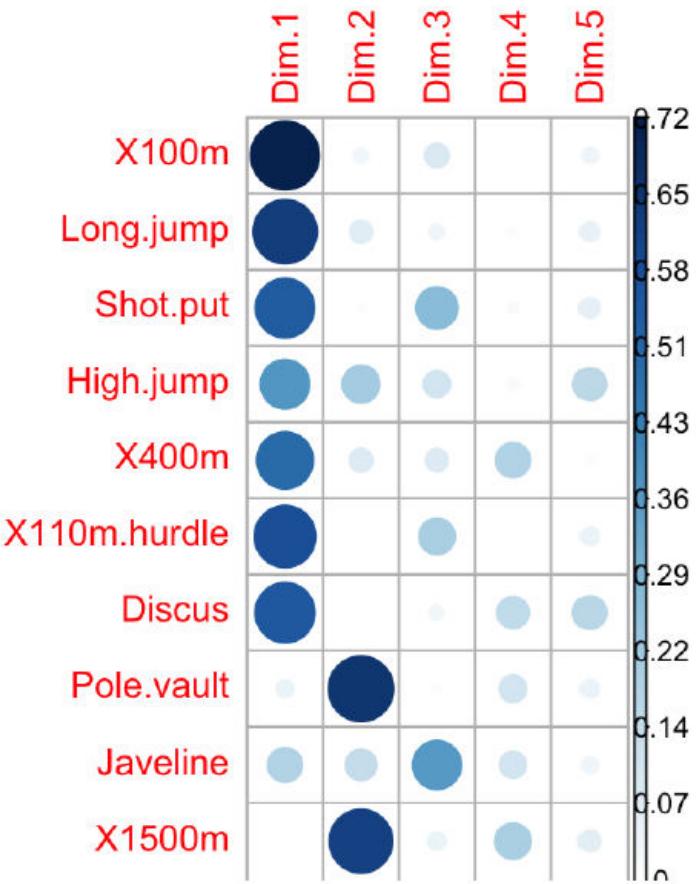
DATA ANALYTICS

PCA using R (factoMineR , factoextra)

- Correlation circle



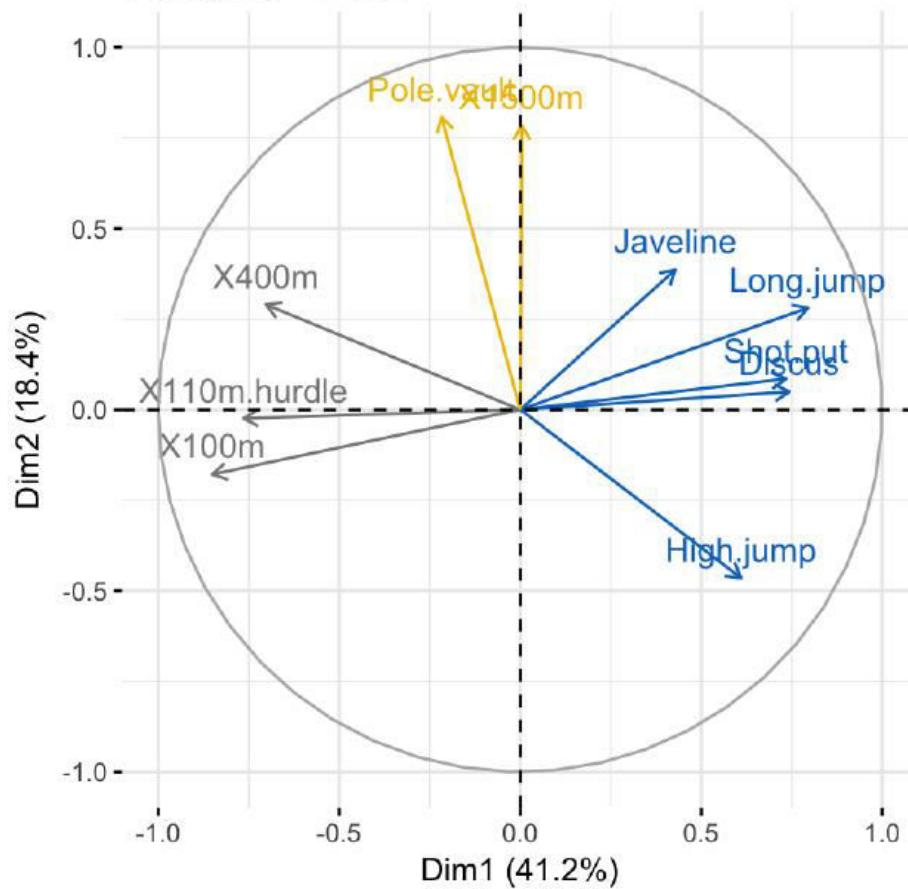
`var.cos2 =
var.coord * var.coord`



PCA using R (factoMineR , factoextra)

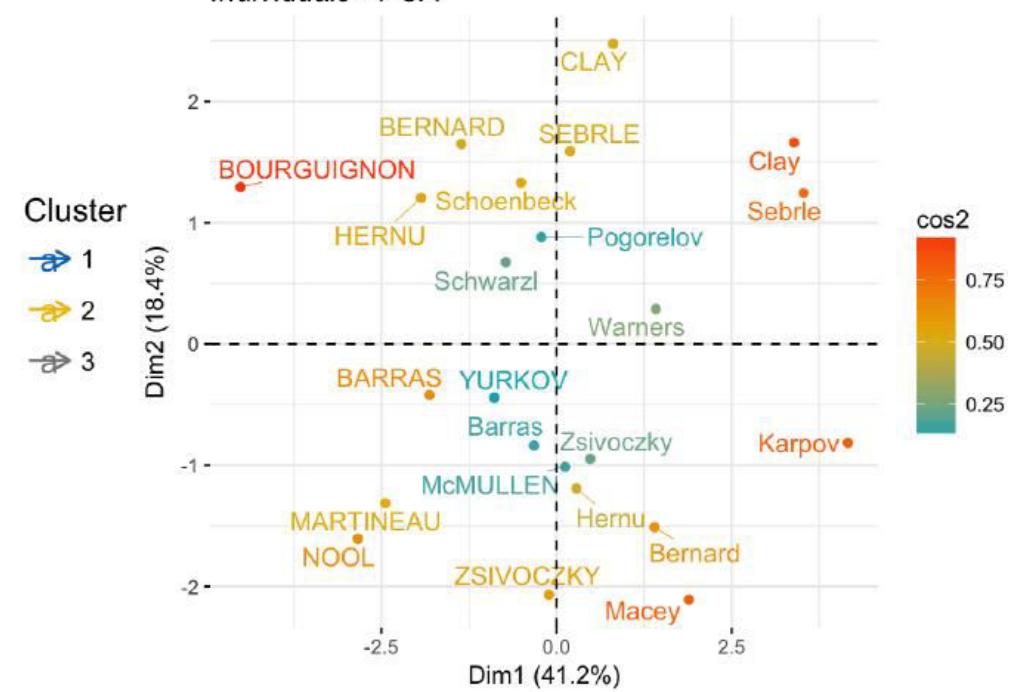
- Which events are similar?

Variables - PCA



- Which athletes are similar?

Individuals - PCA



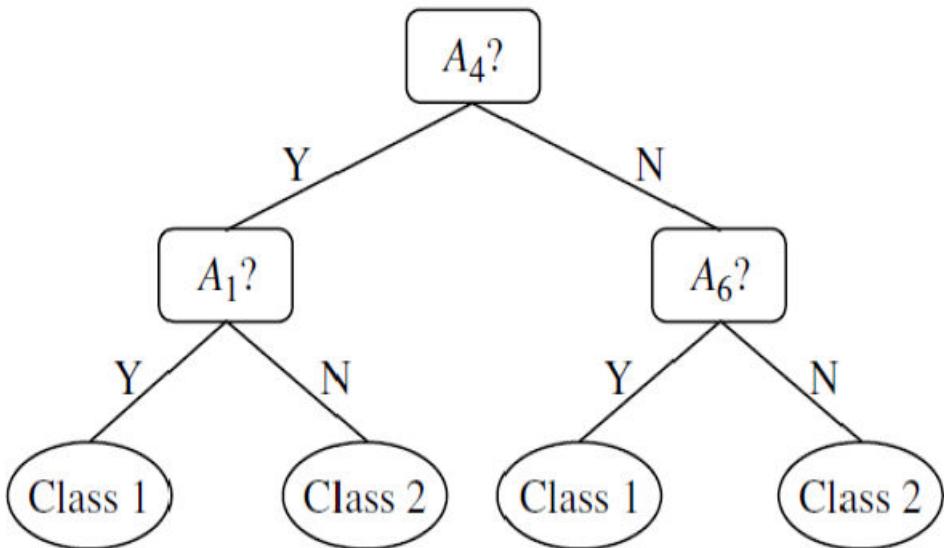
Attribute Subset Selection

- Attribute Subset Selection reduces the dataset size by removing irrelevant or redundant attributes.
- Redundant attributes : Information is contained or can be extracted from other attributes. Example : MRP of a product and the corresponding sales tax paid.
- Irrelevant attributes : They contain no information which is useful for the data analysis task at hand. Example : SRN is irrelevant to predict students' GPA.
- The goal of attribute subset selection is to find the minimum set of attributes such that the resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all attributes.
- Analysis on a reduced set of attributes implies the reduction of the number of attributes appearing in the discovered patterns , thus helping to make the patterns easier to understand.

Heuristic Search in Attribute Subset Selection

- For n attributes , there are 2^n possible subsets. An exhaustive search is infeasible.
- The *best* and the *worst* attributes are determined using tests of statistical significance , which assume that the attributes are independent of each other.
- **Stepwise forward selection** : The procedure starts with an empty set as the reduced set. In each iteration , the best of the original attributes is selected and added to the reduced set.
- **Stepwise backward elimination** : The procedure starts with the full set of attributes. In each iteration , the worst attribute remaining in the set is removed.
- **Combination of forward and backward selection** : In each iteration, the best attribute is selected and the worst attribute is removed.
- **Decision tree induction** : A decision tree is constructed using the given data. All attributes which do not appear in the tree are judged to be irrelevant. Thus the set of attributes appearing in the tree form the reduced subset.

Heuristic Search in Attribute Subset Selection

Forward selection	Backward elimination	Decision tree induction
<p>Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$</p> <p>Initial reduced set: $\{\}$</p> <p>$\Rightarrow \{A_1\}$</p> <p>$\Rightarrow \{A_1, A_4\}$</p> <p>\Rightarrow Reduced attribute set: $\{A_1, A_4, A_6\}$</p>	<p>Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$</p> <p>$\Rightarrow \{A_1, A_3, A_4, A_5, A_6\}$</p> <p>$\Rightarrow \{A_1, A_4, A_5, A_6\}$</p> <p>$\Rightarrow$ Reduced attribute set: $\{A_1, A_4, A_6\}$</p>	<p>Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$</p>  <pre> graph TD A4[A4?] -- Y --> A1[A1?] A4 -- N --> A6[A6?] A1 -- Y --> Class1_1((Class 1)) A1 -- N --> Class2_1((Class 2)) A6 -- Y --> Class1_2((Class 1)) A6 -- N --> Class2_2((Class 2)) </pre> <p>\Rightarrow Reduced attribute set: $\{A_1, A_4, A_6\}$</p>

Attribute Creation – Feature Generation

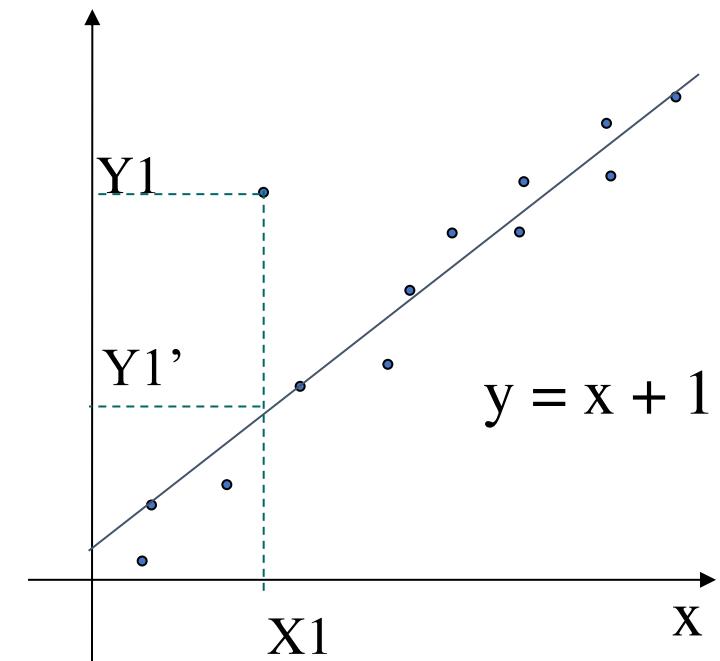
- Creating new features that can capture important information in the dataset more effectively than the original ones.
- For example , an attribute *area* can be added based on the attribute's *height* and *width*. By combining attributes, accuracy can be improved and missing information about the relationships between the attributes can be discovered.
- Three general methodologies
 - Attribute extraction – Domain specific.
 - Mapping data to a new space – Fourier transforms, wavelet transforms and manifold approaches
 - Attribute construction
 - Combining features
 - Data Discretization

Numerosity Reduction

- Reduce data volumes by choosing alternative, smaller forms of data representation.
- **Parametric methods**
 - Assume the data fits some model. Estimate the model parameters and store only the parameters , discarding the data (except the possible outliers).
 - Examples : linear regression , multiple regression , log-linear model
- **Non-Parametric methods**
 - Do not assume models.
 - Examples : histograms, clustering ,sampling

Regression Analysis

- A collective name for techniques for the modeling and analysis of numerical data consisting values of a *dependent variable* and of one or more *independent variables*.
- The parameters are estimated so as to give a *best fit* of the data. The best fit is often evaluated using the least squares method.
- Regression analysis is used for prediction (including forecasting of time series data) , inference, hypothesis testing and modeling of causal relationships.

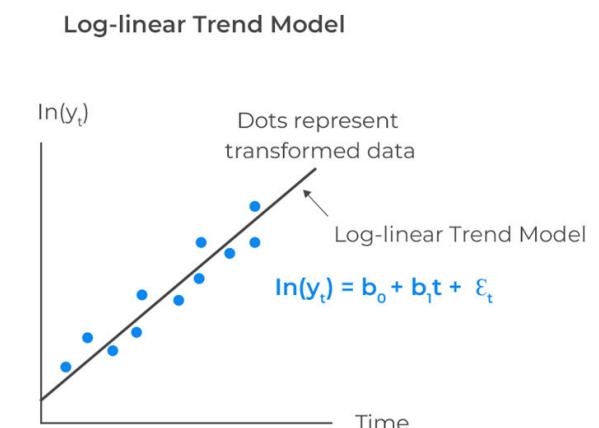
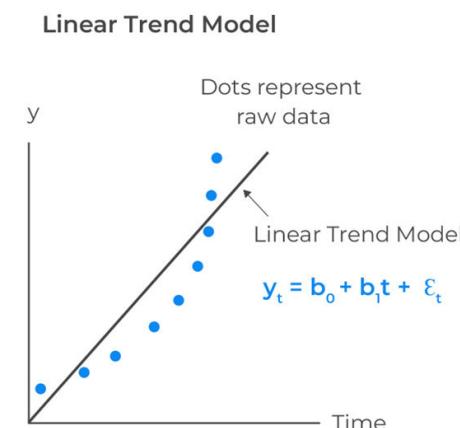


Log-linear model

- A log-linear model is a mathematical model that takes the form of a function whose logarithm equals a linear combination of the parameters of the model.
- It approximates discrete multidimensional probability distributions for a set of discretized attributes based on smaller subset of dimensional combinations.
- It is useful for dimensionality reduction and data smoothening.



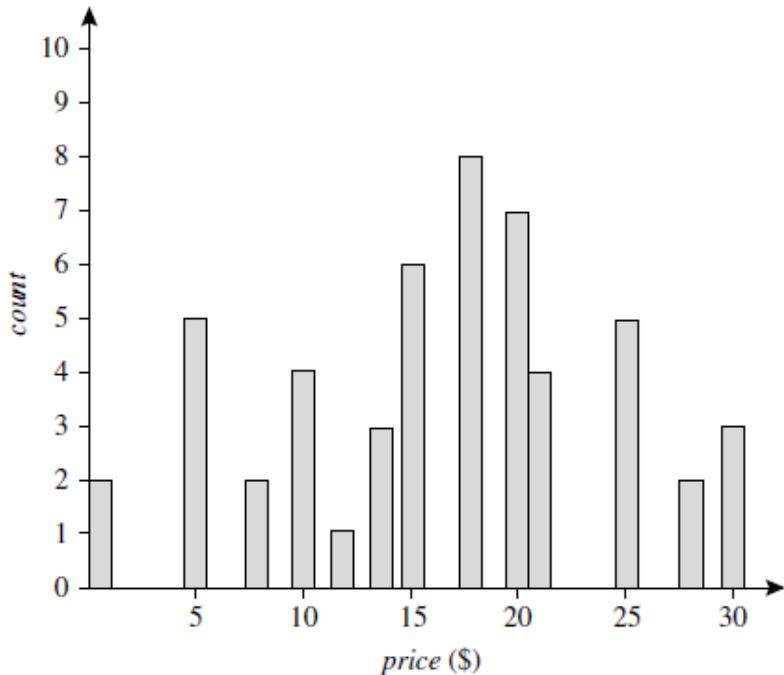
Linear vs. Log-linear Trend Models



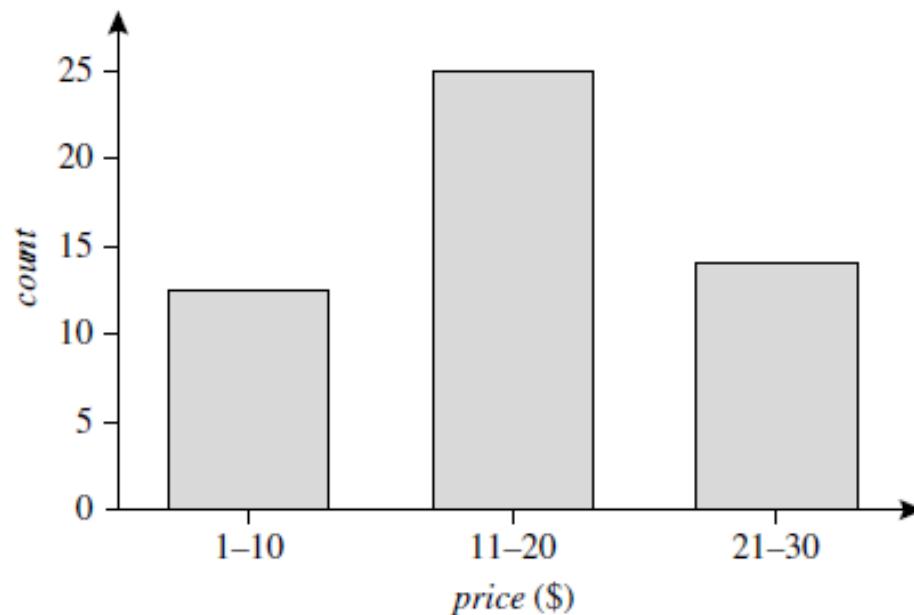
- A histogram for an attribute partitions data distribution into disjoint subsets referred to as bins or buckets.
- If each bucket represents only a single attribute value , it is called singleton bucket.
- Partitioning rules :
 - Equal-width : equal bucket range
 - Equal-frequency : equal depth
- The following data are a list of AllElectronics prices for commonly sold items in \$. The numbers have been sorted: 1, 1, 5, 5, 5, 5, 5, 8, 8, 10, 10, 10, 10, 12, 14, 14, 14, 15, 15, 15, 15, 15, 15, 18, 18, 18, 18, 18, 18, 18, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 28, 28, 30, 30, 30.

Draw a histogram using singleton buckets and equal-width bin of 10\$.

Histogram Analysis



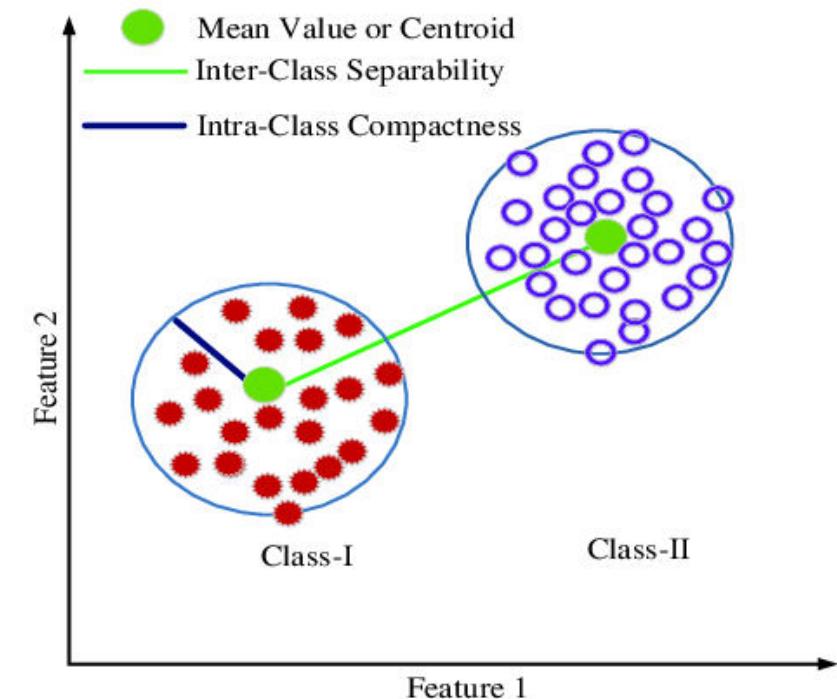
Using singleton buckets



Using a width of 10\$

Clustering

- Records are partitioned into clusters where records within the cluster are similar to one another and dissimilar to the ones in another cluster.
- Cluster representation of data is used to replace the actual data.
- The cluster representation can be the centroid and the diameter(intra-cluster distance) of the cluster.
- It is very effective if the data is *clustered* but not if it is *smeared*.
- Can use hierarchical clustering and be stored in multi-dimensional index trees.

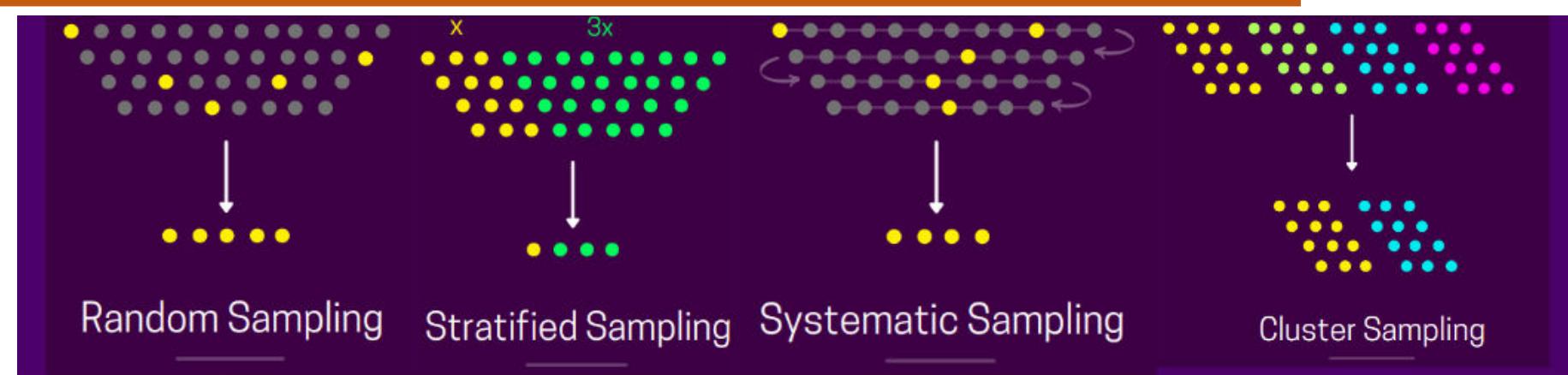


https://www.researchgate.net/figure/The-concept-of-intra-class-compactness-and-inter-class-separability-in-a-two-dimension_fig3_325095062

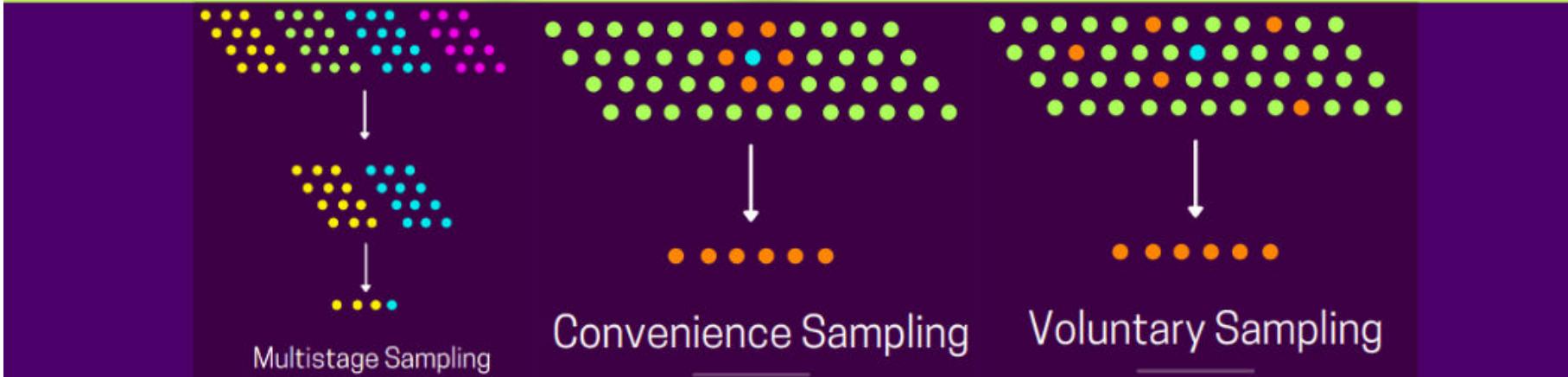
Sampling

- Sampling can be used as a data reduction technique as it allows a large dataset to be represented by a much smaller random subset.
- An advantage of sampling for data reduction is that the cost of obtaining a sample is proportional to the size of the sample.
- In the context of data reduction , sampling is most commonly used to estimate an answer to an aggregate query.
- Using central limit theorem (recall from Statistics for Data Science!!) , it is possible to determine a sufficient sample size for estimating a given function within a specified degree of error.
- Important to remember :
 - Choose a representative subset of the data.
 - Simple random sampling might have a poor performance in presence of a skew.
 - Develop adaptive sampling methods like stratified sampling.

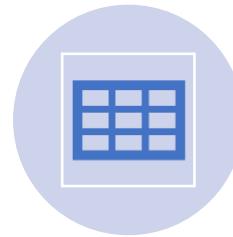
Types of Sampling methods



Sampling Methods



Data Cube Aggregation



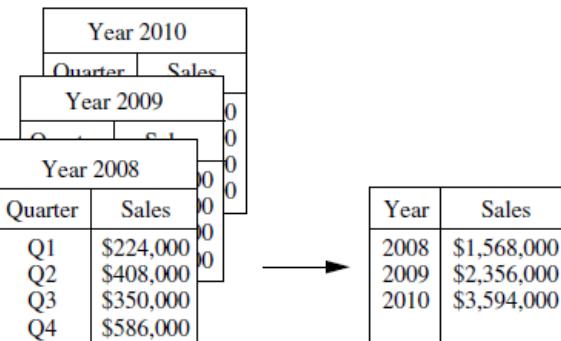
IMAGINE YOU HAVE TO PERFORM AN ANALYSIS ON YEARLY SALES AT DUNDER MIFFLIN PAPER COMPANY.

THE DATA YOU RECEIVE HAS SALES PER QUARTER FROM THE YEAR 2008 TO 2010.

SINCE YOU CARE ABOUT ANNUAL METRICS, THE DATA CAN BE AGGREGATED SO THAT THE RESULTING DATA SUMMARIZES THE ANNUAL SALES RATHER THAN QUARTERLY SALES.

THE RESULTING DATASET IS SMALLER IN VOLUME WITHOUT A LOSS OF INFORMATION NECESSARY TO THE TASK AT HAND!

USUALLY DATA CUBES ARE USED TO STORE MULTIDIMENSIONAL AGGREGATED INFORMATION.



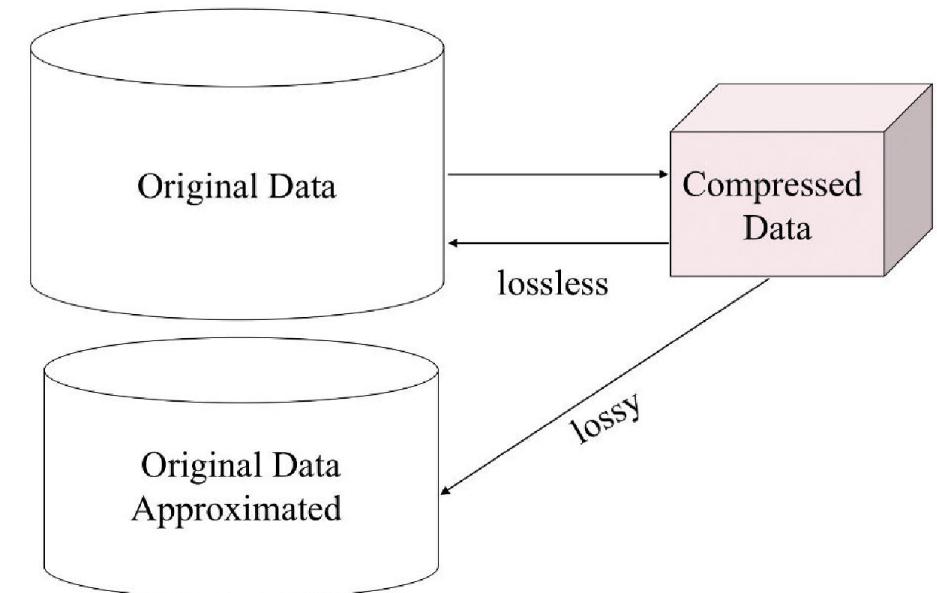
The diagram illustrates the process of data cube aggregation. On the left, there is a complex, multi-layered table representing quarterly sales data for three years: 2008, 2009, and 2010. This table has multiple levels of headers for Year, Quarter, and Sales. An arrow points from this detailed table to a much simpler, aggregated table on the right. The aggregated table has two columns: 'Year' and 'Sales'. It contains three rows, one for each year, summarizing the total sales for that year: 2008 (\$1,568,000), 2009 (\$2,356,000), and 2010 (\$3,594,000).

Year 2010		
Quarter	Sales	
Q1	\$224,000	
Q2	\$408,000	
Q3	\$350,000	
Q4	\$586,000	

Year	Sales
2008	\$1,568,000
2009	\$2,356,000
2010	\$3,594,000

Data Compression

- Transformations are applied so as to obtain a reduced or *compressed* representation of the original data.
- If the original data can be reconstructed from the compressed data without any information loss, the data reduction is called lossless. But if we can reconstruct only an approximation of the original data , then the data reduction is called lossy.
- There are several lossless algorithms for string compression , however they allow only limited data manipulation.
- Dimensionality reduction and numerosity reduction can also be considered as data compression.



Test your understanding!

- What is the first step in data integration?

Understanding the metadata

- Is PCA lossy or lossless?

Lossy

- Amongst **PCA** and **Attribute subset selection**, which data reduction method has more interpretability?

Attribute Subset Selection

Test your understanding!

- For an Electrocardiography (ECG) wave which is a better transform: Fourier or wavelet?

Wavelet

Because ECG's have signals have short intervals of characteristic oscillation and Fourier transforms can only capture frequencies that persist over an entire signal which is not suitable here.

- ----- is a nonzero vector that stays parallel after matrix multiplication

Eigen Vectors

References

- [Data Mining: Concepts and Techniques](#) by Jiawei Han, Micheline Kamber and Jian Pei, The Morgan Kaufmann Series in Data Management Systems, 3rd Edition Chapter : 3.3 – 3.4
- <https://www.st-andrews.ac.uk/~wjh/dataview/tutorials/dwt.html>
- <https://towardsdatascience.com/the-wavelet-transform-e9cfa85d7b34>
- https://www.cs.unm.edu/~mueen/Teaching/CS_521/Lectures/Lecture2.pdf
- <https://medium.com/analytics-vidhya/understanding-principle-component-analysis-pca-step-by-step-e7a4bb4031d9>



THANK YOU

Dr. Gowri Srinivasa

Professor, Department of Computer Science
and Engineering, PES University, Bengaluru

Email: gsrinivasa@pes.edu



DATA ANALYTICS

UE21CS342AA2

UNIT-1

**Lecture 7 : Data Preprocessing –
Transformations and Discretization**

Gowri Srinivasa

Department of Computer Science and Engineering



Data Analytics

Unit 1

Lecture 7 : Data Preprocessing – Transformations and Discretization

Slides excerpted from: Data Mining : Concepts and Techniques by Han, Kamber and Pei, 3rd Edition

Gowri Srinivasa

Department of Computer Science and Engineering

Slides collated by:

Nishanth M S, CSE 2023, PES University

nishanthmsathish.23@gmail.com

Harshitha Srikanth, VII CSE, PES University

harshithasrikanth13@gmail.com

With grateful thanks for contribution of slides to:
Dr. Mamatha H R, Professor at the Department of CSE, PESU

Data Transformation

In data transformation, the data are transformed or consolidated into forms appropriate

for mining. Strategies for data transformation include the following:

A function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values.

Methods

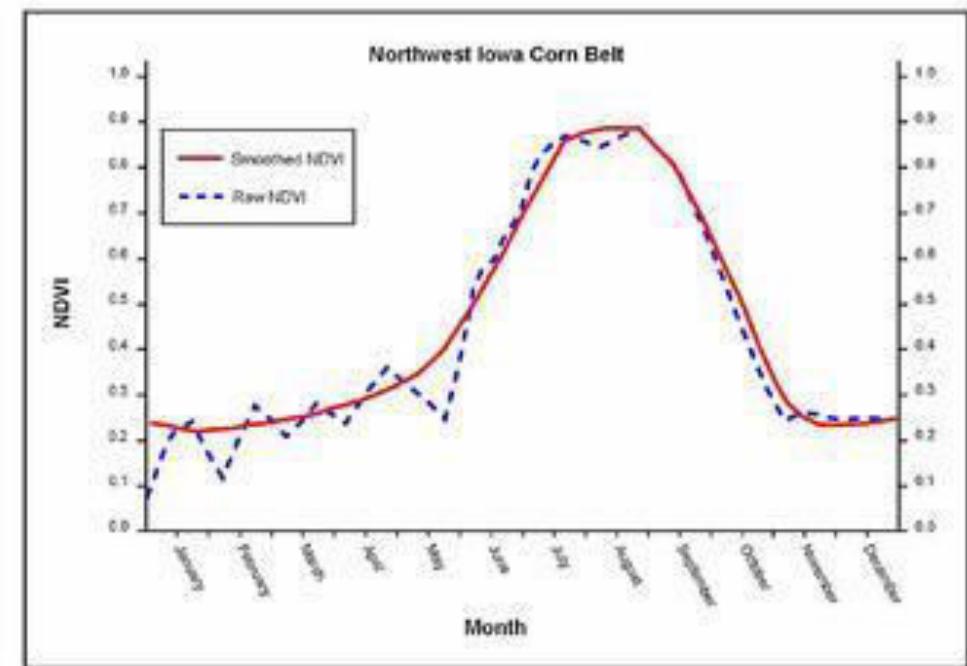
- Smoothing
- Attribute construction
- Aggregation
- Normalization
- Discretization
- Concept hierarchy generation

Smoothing

Removal of noise in data.

Techniques:

- Binning
- Regression
- Clustering
- Simple average – Time series data
- Weighted average – Time series data
- Exponential smoothening – Time series data
- Gaussian filter - Image





Normalization

Need for normalization :

- The unit of measurement can affect data analysis.
- For example , changing the unit of measurement of height from inches to cm can lead to different results.
- An attribute in smaller units results in a larger range. This attribute will be given a higher weight(preference) by few distance-based algorithms like KNN(K-nearest neighbors).
- For example , income of a person is generally a few orders higher than their height.
- To overcome this , data is generally transformed to a range such as [-1,1] or [0,1].

Sample Data:

House	Area (in sq.ft.)	Number of Bedrooms
A	2000	3
B	2050	4
C	2500	3
D	2200	2

Normalization

Normalization by decimal scaling normalizes by moving the decimal point of values of attribute A . The number of decimal points moved depends on the maximum absolute value of A . A value, v_i , of A is normalized to v'_i by computing

$$v'_i = \frac{v_i}{10^j}, \quad (3.12)$$

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

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

Normalization

- **Min-Max Normalization :** Performs a linear transformation. It transforms the values from $[min_A, max_A]$ to $[new_min_A, new_max_A]$. A value v is transformed by

$$v' = \frac{v - min_A}{max_A - min_A} (new_max_A - new_min_A) + new_min_A$$

- Example : Let the income range \$12,000 to \$98,000 be normalized to [0.0,1.0]. Find out the mapping for \$73,600

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

- This preserves the relationship among the original data values. It will encounter an “out-of-bounds” error if an input which is outside the range of original data is provided.

Normalization

- **Z-score normalization** : The values for an attribute A are normalized based on the mean and standard deviation of A. A value v can be normalized by

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

Where μ_A is the mean and σ_A the standard deviation.

- Example : Let $\mu = 54,000$ and $\sigma = 16,000$. Then z-score of 73,600 is

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

- This method of normalization is useful when the actual minimum or maximum value of A is unknown or there are outliers in A.
- A variation of this is obtained by using Mean Absolute Deviation (MAD) instead of standard deviation. It is less susceptible to outliers.

Discretization

- Divides the range of a continuous attribute into intervals.
- Interval labels are then used to replace the actual data values.
- Data size can be reduced by discretization.
- If the discretized process uses class information , it is called *supervised discretization* else it is called *unsupervised discretization*.
- **Top-down discretization** : Process starts by finding one or few points (called split or cut points) to split the entire attribute range and then repeat this process recursively on the resulting intervals.
- **Bottom-up discretization** : Also called as *merging* , starts by considering all the continuous values as potential splits. It removes few split points by merging neighborhood values to form intervals. This process is recursively applied to the resulting intervals.

Data discretization methods

- Binning :
 - Top-down split , unsupervised
- Histogram analysis :
 - Top-down split , unsupervised
- Clustering analysis :
 - Unsupervised , top-down split or bottom-up merge
- Decision-tree analysis :
 - Supervised , top-down split
- Correlation analysis :
 - Unsupervised , bottom-up merge

Note : All these methods can be applied recursively.

Binning

- Equal-width (distance) partitioning
 - Divides the range into N intervals of equal size.
 - The width of the interval is $w = (\text{Maximum} - \text{Minimum})/N$.
 - Is susceptible to outliers and skewed data.
- Equal-depth (frequency) partitioning
 - Divides the range into N intervals, each containing approximately same number of samples.
 - Ensures good data scaling but managing categorical attributes can get tricky.

Binning - Example

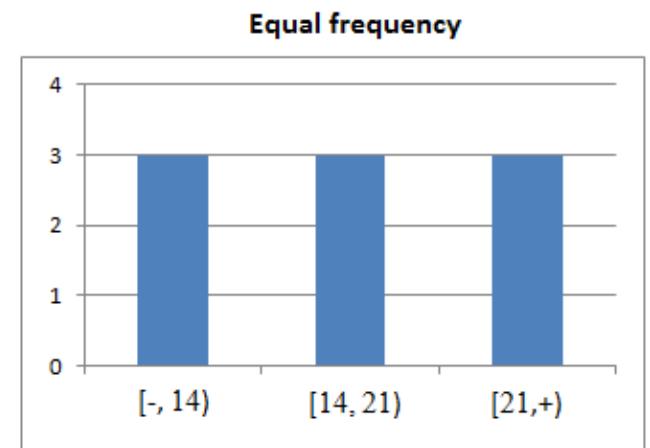
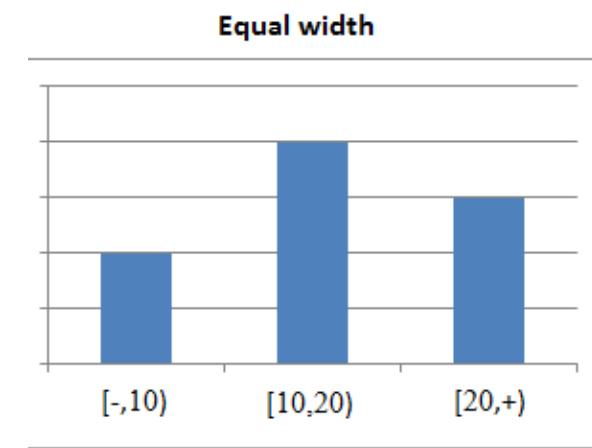
- Data : 0, 4, 12, 16, 16, 18, 24, 26, 28

- Equal width

- Bin 1: 0, 4 [-,10)
- Bin 2: 12, 16, 16, 18 [10,20)
- Bin 3: 24, 26, 28 [20,+)

- Equal frequency

- Bin 1: 0, 4, 12 [-, 14)
- Bin 2: 16, 16, 18 [14, 21)
- Bin 3: 24, 26, 28 [21,+)



Data Smoothing with Binning

Sorted data for price (in \$) : 4,8,9,15,21,21,24,25,26,28,29,34

Partition into equal-depth (frequency) bins

Bin-1 : 4,8,9,15

Bin-2 : 21,21,24,25

Bin-3 : 26,28,29,34

Data Smoothing with Binning

- Smoothing by **bin means** :

Bin-1 : 9,9,9,9

Bin-2 : 23,23,23,23

Bin-3 : 29,29,29,29

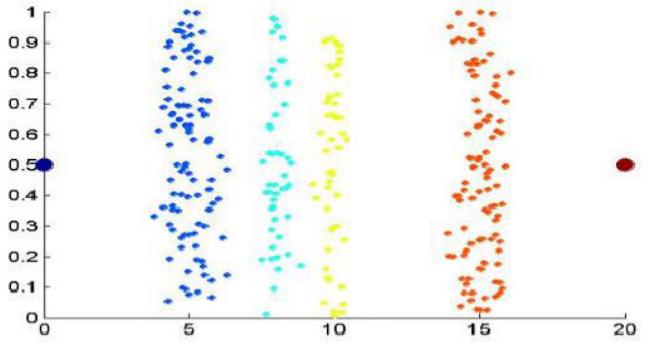
- Smoothing by **bin boundaries** :

Bin-1 : 4,4,4,15

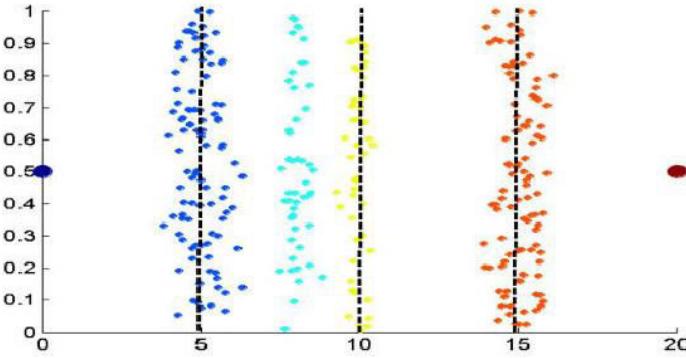
Bin-2 : 21,21,25,25

Bin-3 : 26,26,26,34

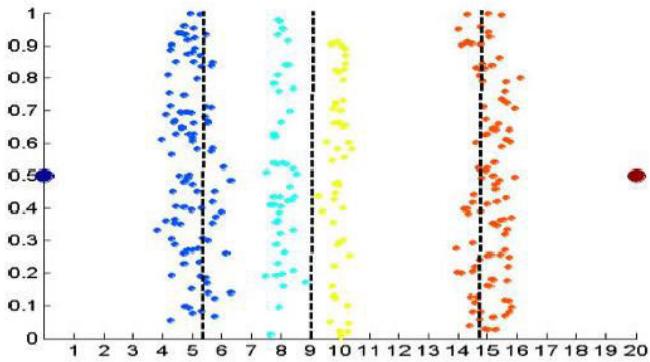
Binning vs Clustering (Unsupervised Discretization)



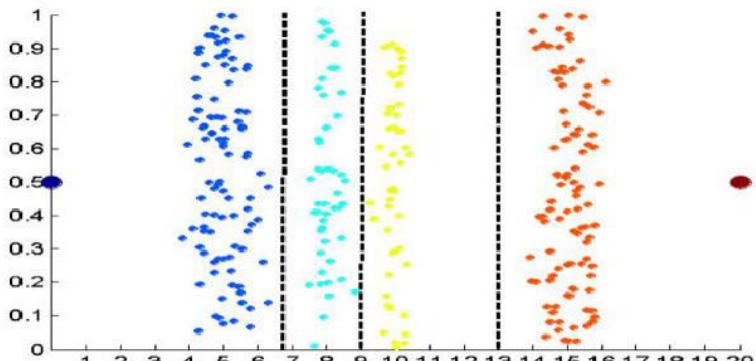
Original Data



Equal Width(binning)



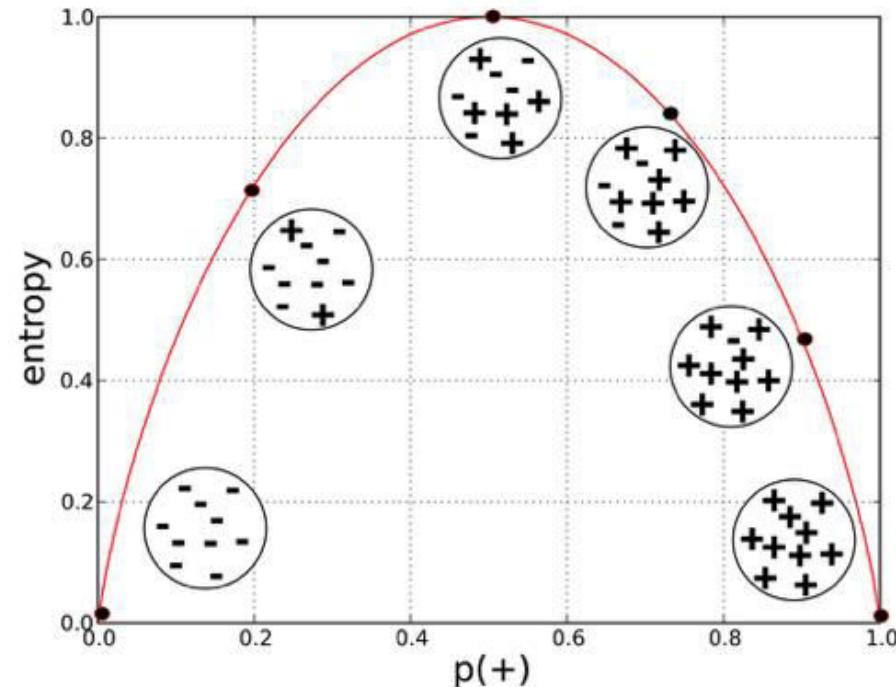
Equal frequency (binning)



K-means clustering leads to better results

Discretization by Classification

- Supervised : Class labels are used in determining the split point.
- It is a top-down discretization where recursive split is applied.
- Example : Decision tree analysis.
- Entropy is used to determine the split point. Lower the entropy , better the split. It is given by $E(S) = \sum_{i=1}^c - p_i \log_2 p_i$



Discretization by Correlation Analysis

- **Supervised : Class labels are used in determining the split point.**
- Example , Chi-merge: χ^2 -based discretization. It is a bottom-up merge.
- Initially each distinct value of the attribute is considered to be one interval.
- χ^2 tests are performed for every pair of adjacent intervals.
- **Adjacent intervals with the least χ^2 values are merged as it indicates similar class distributions.**
- This merging process proceeds recursively until a pre-defined threshold for χ^2 is met.

Discretization by Correlation Analysis

- A low χ^2 value means the observed class distribution closely matches the expected distribution, implying that the class label is nearly independent of whether a data point falls in one interval or the other. In simpler terms, those two intervals are similar in terms of their relationship with the class labels.

- Since these intervals are not contributing much discriminative power separately, they can be merged without losing much information. By merging them, we reduce the complexity of the discretized feature without sacrificing much, if any, of its predictive power.

Sample	F	K
1	1	1
2	3	2
3	7	1
4	8	1
5	9	1
6	11	2
7	23	2
8	37	1
9	39	2
10	45	1
11	46	1
12	59	1

Intervals

{0,2}

$$(1+3)/2=2$$

{2,5}

$$(3+7)/2=5$$

{5,7.5}

{7.5,8.5}

{8.5,10}

{10,17}

{17,30}

{30,38}

{38,42}

{42,45.5}

{45.5,52}

{52,60}

ChiMerge Discretization Example



- Sort and order the attributes that you want to group (in this example attribute F).

- Start with having every unique value in the attribute be in its own interval.

Discretization by Correlation Analysis - Example

Sample	F	K
1	1	1
2	3	2
3	7	1
4	8	1
5	9	1
6	11	2
7	23	2
8	37	1
9	39	2
10	45	1
11	46	1
12	59	1

ChiMerge Discretization
Example

- Begin calculating the Chi Square test on every interval

Sample	K=1	K=2	
2	0	1	1
3	1	0	1
total	1	1	2

Sample	K=1	K=2	
3	1	0	1
4	1	0	1
total	2	0	2

Sample	K=1	K=2	
2	0	1	1
3	1	0	1
total	1	1	2

$$E_{11} = (1/2)*1 = .05$$

$$E_{12} = (1/2)*1 = .05$$

$$E_{21} = (1/2)*1 = .05$$

$$E_{22} = (1/2)*1 = .05$$

$$\chi^2 = (0-.5)^2/.5 + (0-.5)^2/.5 + (0-.5)^2/.5 + (0-.5)^2/.5 = 2$$

Sample	K=1	K=2	
3	1	0	1
4	1	0	1
total	2	0	2

$$E_{11} = (1/2)*2 = 1$$

$$E_{12} = (0/2)*2 = 0$$

$$E_{21} = (1/2)*2 = 1$$

$$E_{22} = (0/2)*2 = 0$$

$$\chi^2 = (1-1)^2/1 + (0-0)^2/0 + (1-1)^2/1 + (0-0)^2/0 = 0$$

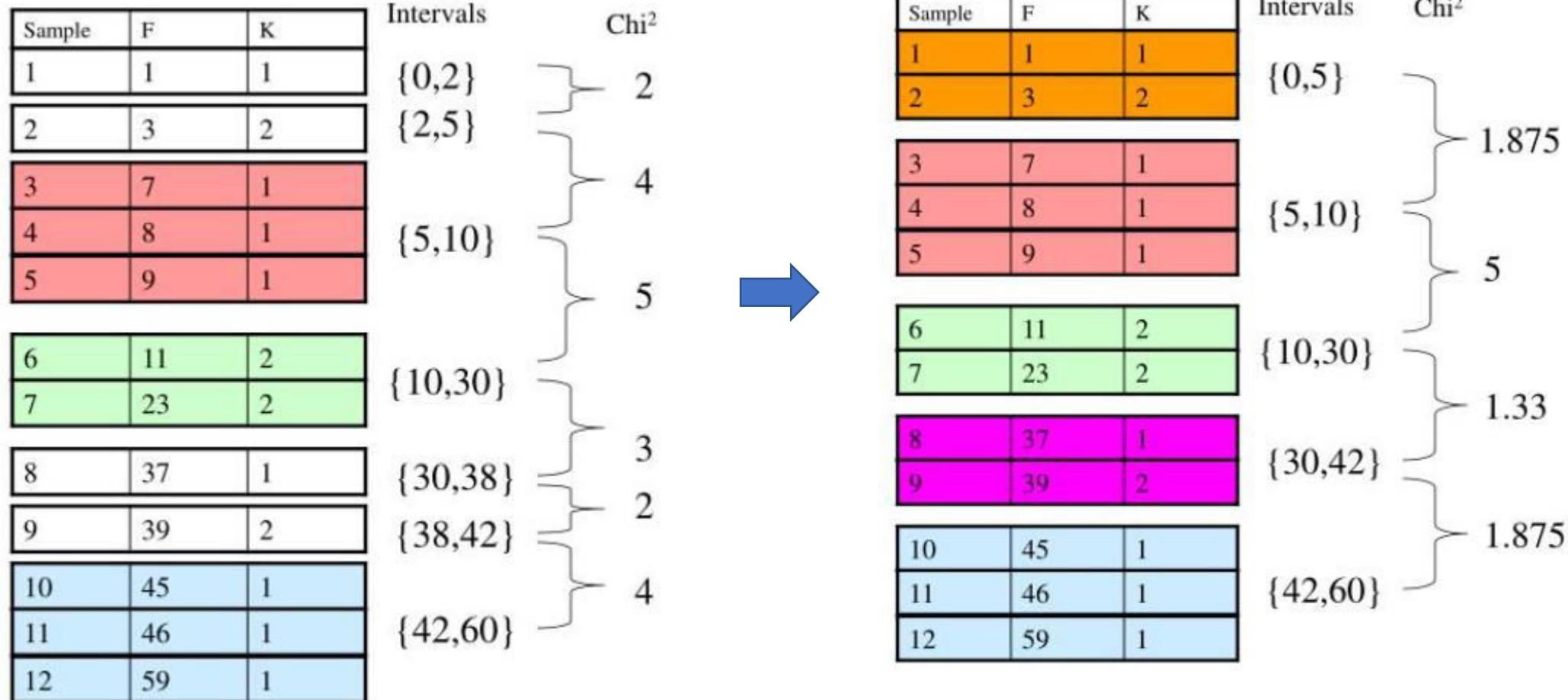
Threshold .1 with df=1 from Chi square distribution chart merge if
 $\chi^2 < 2.7024$

Sample	F	K	Intervals	Chi ²
1	1	1	{0,2}	2
2	3	2	{2,5}	2
3	7	1	{5,7.5}	0
4	8	1	{7.5,8.5}	0
5	9	1	{8.5,10}	2
6	11	2	{10,17}	0
7	23	2	{17,30}	2
8	37	1	{30,38}	2
9	39	2	{38,42}	2
10	45	1	{42,45.5}	0
11	46	1	{45.5,52}	0
12	59	1	{52,60}	0

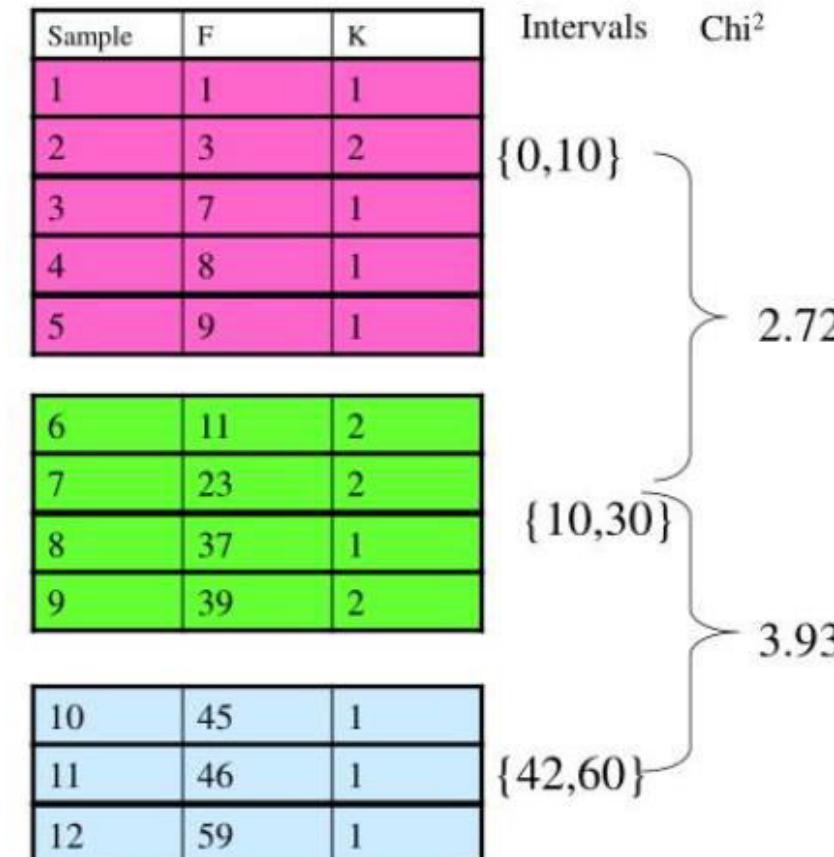
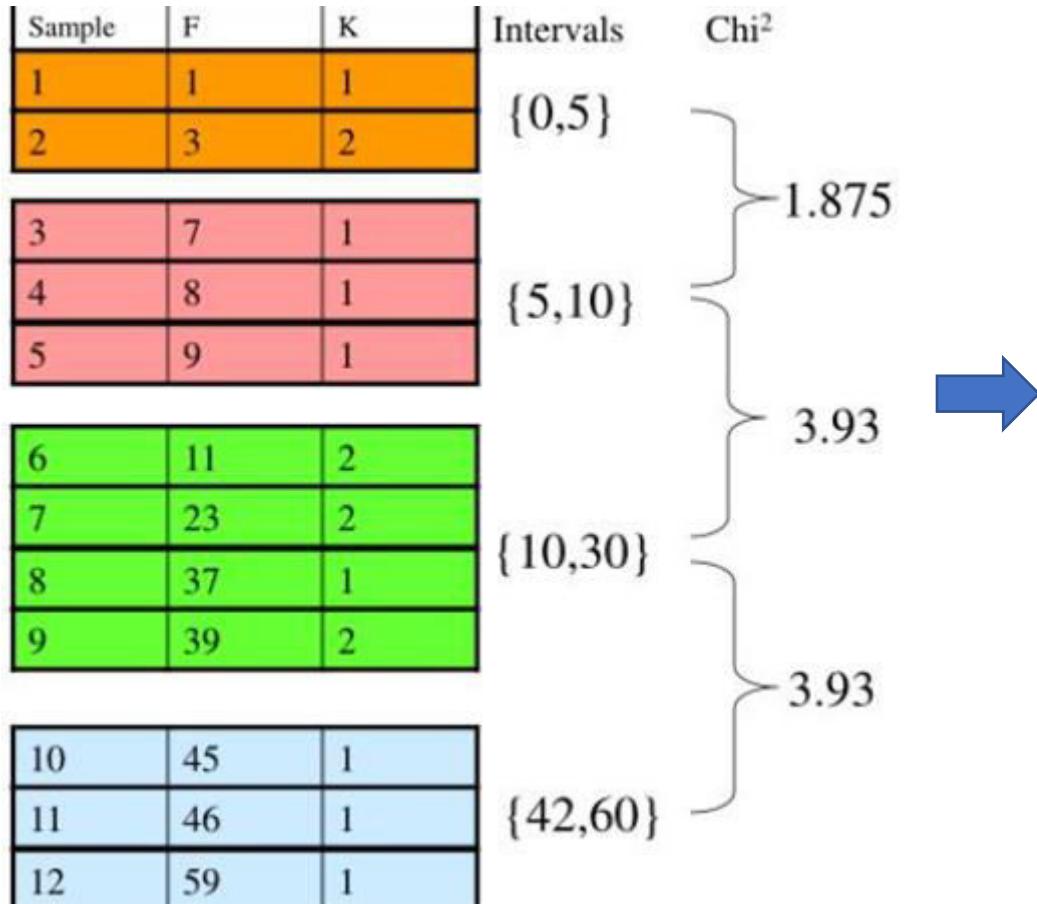
- Calculate all the Chi Square value for all intervals

- Merge the intervals with the smallest Chi values

Discretization by Correlation Analysis - Example



Discretization by Correlation Analysis - Example



- There are no more intervals that can satisfy the Chi Square test.

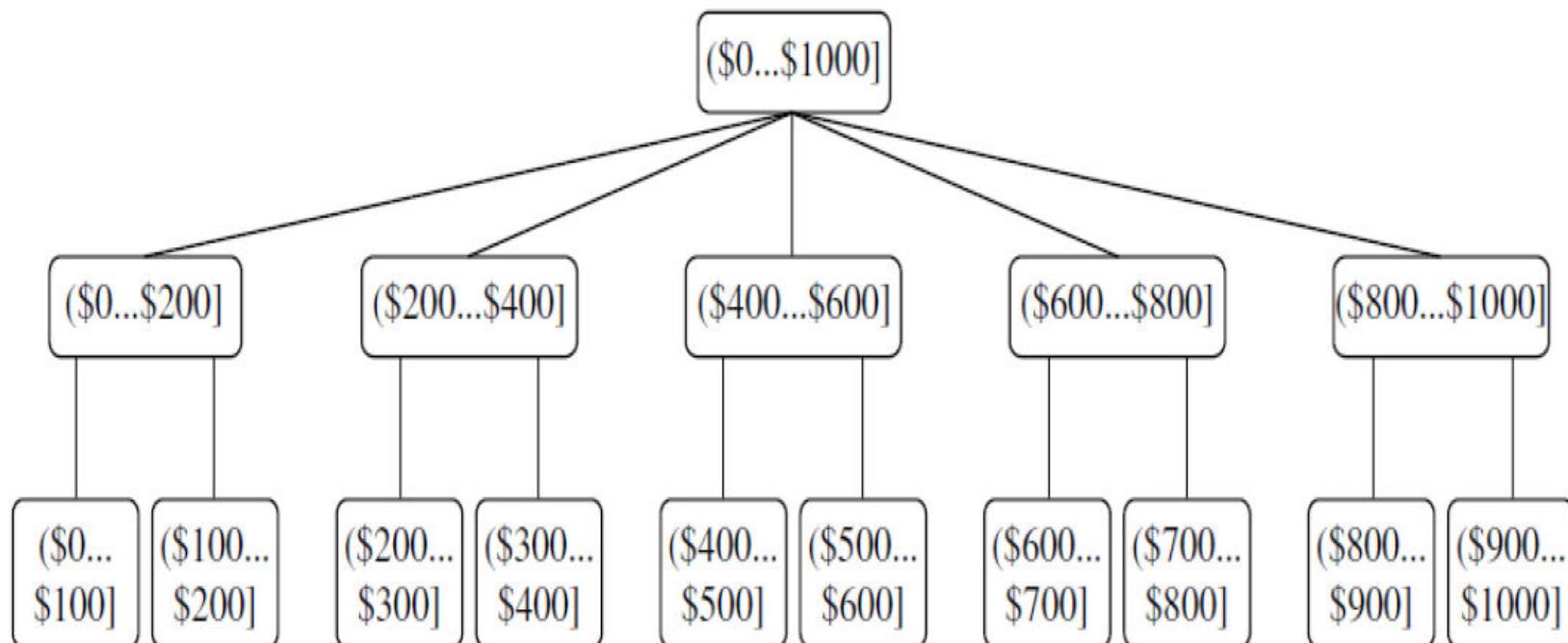
Concept Hierarchy Generation

- Concept hierarchy organizes concepts (Attribute values) hierarchically by representing a series of mappings from a set of low-level concepts to a high-level , generalized concepts.
- It facilitates drilling and rolling in data warehouses to view data in multiple granularity.
- Method : Recursively reduce the data by collecting and replacing low level concepts (such as numeric values for age) by higher level concepts (such as kids, teenagers, adults, senior citizens).
- They can be explicitly specified by domain experts and/or data warehouse designers.

Concept Hierarchy Generation for Numeric Data

Discretization methods discussed till now can be used for numeric data.

Example:



Concept Hierarchy Generation for Nominal Data

1) Specification of a partial ordering of attributes explicitly at the schema level

- A user or an expert defines a concept hierarchy by specifying a partial or a total ordering of attributes at the schema level.
- For example , suppose a relational database contains the attributes *street*, *city*, *state* and *country* . *Location* dimension of the data warehouse may contain the same attributes.
- A hierarchy can be defined by specifying the total ordering among these attributes at the schema level

street < city < state < country

Concept Hierarchy Generation for Nominal Data

2) Specification of a portion of hierarchy by explicit data grouping

- A portion of the concept hierarchy is manually defined.
- In a large database , it is unrealistic to define the entire concept hierarchy by explicit value enumeration.
- However , we can easily specify explicit groupings for a small portion of intermediate-level data.
- For example , after specifying state and country form a hierarchy at the schema level , a user can define few intermediate levels manually

{Karnataka , Tamil Nadu, Kerala, Andhra Pradesh, Telangana} < South India

Concept Hierarchy Generation for Nominal Data

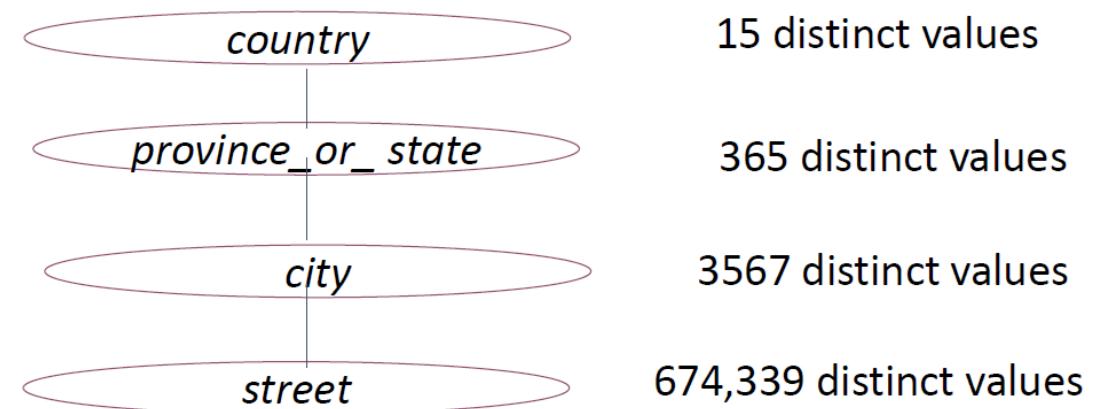
3) Specification of only a partial set of attributes.

- At times , a user can have a vague idea about what should be included in the hierarchy.
- The user may have included only a small subset of the relevant attributes in the hierarchy specification.
- For example, instead of including all the hierarchically relevant attributes for *location* , the user might have specified only *street* and *city*.
- To handle this , embed data semantics into the database schema. Hence one attribute will trigger a whole group of linked attributes to be added to the hierarchy. For example , when *city* is added , it would automatically include *state* and *country* as they are semantically related.

Concept Hierarchy Generation for Nominal Data

4) Automatic generation of hierarchies by analysis of distinct values per attribute

- Few hierarchies can be automatically generated based on the analysis of the number of distinct values per attribute in the dataset.
- The attribute with the most distinct values is placed at the lowest level of the hierarchy.



Note : This method is not foolproof. For example, a time dimension in a database might contain 20 distinct *years* , 12 distinct *months* and 7 distinct *days of the week*. However , this doesn't suggest that the time hierarchy should be *year<month<days of the week* .

Test your understanding!

- Which method of normalization must one choose if they are dealing with a lot of outliers and don't know the range of their data?

Solution

Z-Score Normalization

- Which split-point is preferred for discretization using decision trees?

Solution

A Split point which results in least entropy.

- Which normalization method strictly works in the range of input data?

Solution

Min-Max Normalization

Test your understanding!

- Consider a set of Unsorted data for price in dollars
8 16, 9, 15, 21, 21, 24, 30, 26, 27, 30, 34

- Smooth the data by equal frequency bins
- On the results of part (1) apply smoothing by bin means

Solution:

1)

Bin 1: 8, 9, 15, 16

Bin 2: 21, 21, 24, 26,

Bin 3: 27, 30, 30, 34

For Bin 1:

$$(8 + 9 + 15 + 16 / 4) = 12$$

(4 indicating the total values like 8, 9, 15, 16)

Bin 1 = 12, 12, 12, 12

For Bin 2:

$$(21 + 21 + 24 + 26 / 4) = 23$$

Bin 2 = 23, 23, 23, 23

For Bin 3:

$$(27 + 30 + 30 + 34 / 4) = 30$$

Bin 3 = 30, 30, 30, 30

2)

References

- [**Data Mining: Concepts and Techniques**](#) by Jiawei Han, Micheline Kamber and Jian Pei, The Morgan Kaufmann Series in Data Management Systems, 3rd Edition Chapter 3.5
- <https://t4tutorials.com/binning-methods-for-data-smoothing-in-data-mining/>



THANK YOU

Dr. Gowri Srinivasa

Professor, Department of Computer Science
and Engineering, PES University, Bengaluru

Email: gsrinivasa@pes.edu





PES
UNIVERSITY

DATA ANALYTICS

UE21CS342AA2

UNIT-1

Lecture 8 : Analysis of Variance - 1

Gowri Srinivasa

Department of Computer Science and Engineering

Data Analytics

Unit 1

Lecture 8 : Analysis of Variance - 1

Slides excerpted from: U. Dinesh Kumar,
“Business Analytics”, Wiley, 2nd Edition 2022

Gowri Srinivasa

Department of Computer Science and Engineering

Slides collated by:

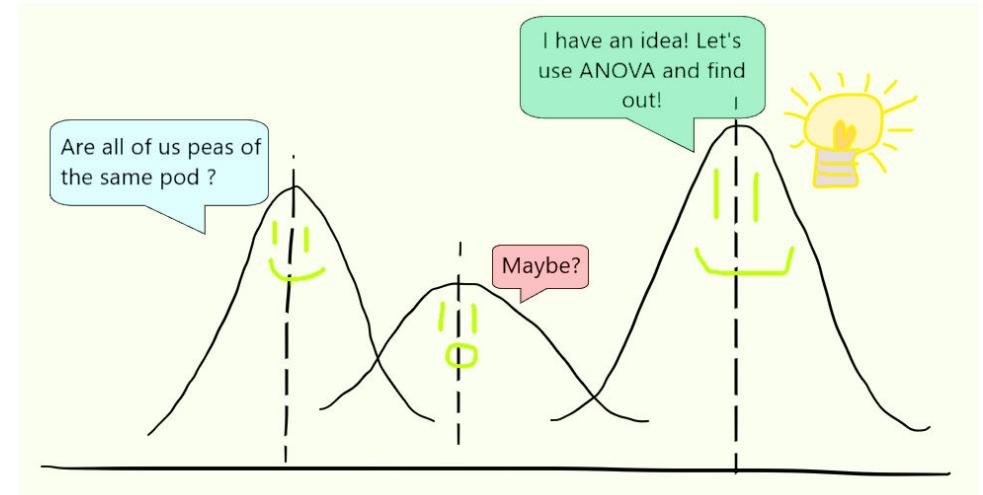
Nishanth M S, CSE 2023, PES University
nishanthmsathish.23@gmail.com

Harshitha Srikanth, VII CSE, PES University
harshithasrikanth13@gmail.com

With grateful thanks for contribution of slides to:
Dr. Mamatha H R, Professor at the Department of CSE, PESU



- Analysis of Variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other.
- **ANOVA checks the impact of one or more factors by comparing the means of different samples.**
- For example , assume there are 3 classes. For a given exam , you want to find out if the marks of the student (dependent variable) depends on the class they study in. So , we are trying to find out if *class they study in* is a **factor**.



<https://www.geeksforgeeks.org/one-way-anova/>

To understand whether the factor (different levels of factor) has any statistical significance on the population parameter ,we compare to models

1) Means Model : The reduced model essentially posits that there's no difference among the means of different levels of the factor, and any observed differences are purely due to random error.

- It is given by $Y_{ij} = \mu + \varepsilon_{ij}$
- Y_{ij} is the value of the outcome variable of j^{th} observation for i^{th} factor level, μ is the overall mean value of all observations, ε_{ij} is the error assumed to be a normal distribution with mean 0 and standard deviation σ .
- Model defined in above equation is often called the **reduced model**, in which the mean μ is common for all levels of the factor.

2) Factor Effect Model :

- It is given by $Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$
- μ is the overall mean and τ_i is the effect of factor i (or factor effect). τ_i is the difference between overall mean and the factor level mean.
- A non-zero τ_i implies that a factor has an influence on the value of the outcome variable Y_{ij} .
- Our objective is to verify if the variation is due to factors is different from the variations due to randomness. you're aiming to discern whether the observed differences between groups are meaningful and statistically significant, or if they could just be random noise.
- Model defined in above equation is called **full model**.

Variation Due to Factors: This is the variability in the dependent variable that can be attributed to the different levels or categories of our independent variable (factor). In ANOVA, this is represented by the between-group variability or the sum of squares between groups (SSB).

Variation Due to Randomness: This refers to the inherent variability present within each group or category. It captures the random fluctuations or differences within groups that aren't due to the factor being studied. In ANOVA, this is represented by the within-group variability or the sum of squares within groups (SSW).

In ANOVA, our objective is to verify whether the variation due to treatment is different from the variation due to randomness.

Multiple t-tests for comparing several means

- Consider a retail store who would like to study the impact of different levels of price discounts (factors) on the sales (outcome variable). Let's say they are analyzing the levels of discounts of 0%, 10% and 20%.
- If we had only 2 levels of discount , we could have used a t-test directly to check whether a statistically significant relationship exists between price discount and average sales quantity.
- One option is to use 3 different 2 sample t-test:
 - Test between 0% and 10%
 - Test between 0% and 20%
 - Test between 10% and 20%

Test	Null Hypothesis	Alternative Hypothesis	Significance (α)
A	$H_0: \mu_0 = \mu_{10}$	$H_A: \mu_0 \neq \mu_{10}$	$\alpha = 0.05$
B	$H_0: \mu_0 = \mu_{20}$	$H_A: \mu_0 \neq \mu_{20}$	$\alpha = 0.05$
C	$H_0: \mu_{10} = \mu_{20}$	$H_A: \mu_{10} \neq \mu_{20}$	$\alpha = 0.05$

Multiple t-tests for comparing several means

- Let,

$$P(A) = P(\text{Retain } H_0 \text{ in test A} \mid H_0 \text{ in test A is true})$$

$$P(B) = P(\text{Retain } H_0 \text{ in test B} \mid H_0 \text{ in test B is true})$$

$$P(C) = P(\text{Retain } H_0 \text{ in test C} \mid H_0 \text{ in test C is true})$$

- Note : values of $P(A) = P(B) = P(C) = 1 - \alpha = 1 - 0.05 = 0.95$
- The conditional probability of simultaneously retaining all 3 null hypotheses when they are true is $P(A \cap B \cap C) = 0.95^3 = 0.8573$.
- Now consider the following null hypothesis:

$$H_0: \mu_0 = \mu_{10} = \mu_{20}$$

If we retain the null hypothesis based on the three individual *t*-tests, then the significance or Type I error is not α -value but much higher than α (Lunney, 1969; Siegel, 1990).

α is the significance level of each test, and it's set to 0.05. This is the probability of making a Type I error, which means incorrectly rejecting H_0 when it's actually true.

The need for ANOVA

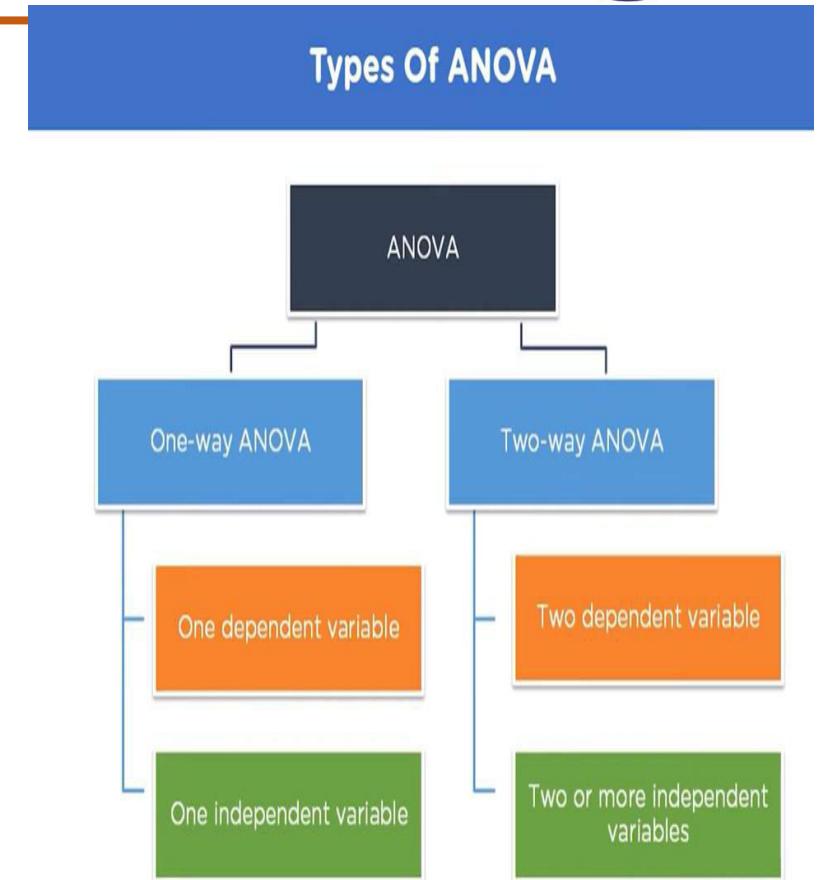
- For the case discussed , if we retain the null hypothesis based on 3 individual tests, then the Type I error is $1 - 0.8573 = 0.1426$.
- When more than 2 groups are involved, checking the population parameter values simultaneously using *t*-tests is inappropriate since the Type I and Type II errors will be estimated incorrectly.
- For this reason, we use analysis of variance (ANOVA) whenever we need to compare 3 or more groups for population parameter values simultaneously.

One-way ANOVA-conditions



One-way ANOVA is appropriate under the following conditions:

- 1) We would like to study the impact of a **single treatment** (also known as factor) at different levels (thus forming different groups) on a continuous **response variable** (or outcome variable). For the example discussed , the variable ‘price discount’ is the treatment (or factor) and 0%, 10%, and 20% price discounts are the different levels (3 levels in this case), different levels of discount is likely to have varying impact on the sales of the product, where sales is the outcome variable. We would like to understand the impact of different levels of price discount on the response variable, sales.



One-way ANOVA-conditions

- 2) In each group, the population **response variable** follows a normal distribution and the sample subjects are chosen using random sampling.(**normality assumption**)
- 3) The population variances for different groups are assumed to be same. That is, variability in the response variable values within different groups is same(homogeneity of variances) asserts that the dispersion of data points within each group is similar across all groups being compared.

Although conditions 2 and 3 are necessary for one-way ANOVA, the model is robust and minor violations of the assumptions may not result in incorrect decision about the null hypothesis.

- Assume that we would like to study the impact of a factor (such as discount) with k levels on a continuous variable (such as sales quantity).
- Then the null and alternative hypotheses for one way ANOVA are given by

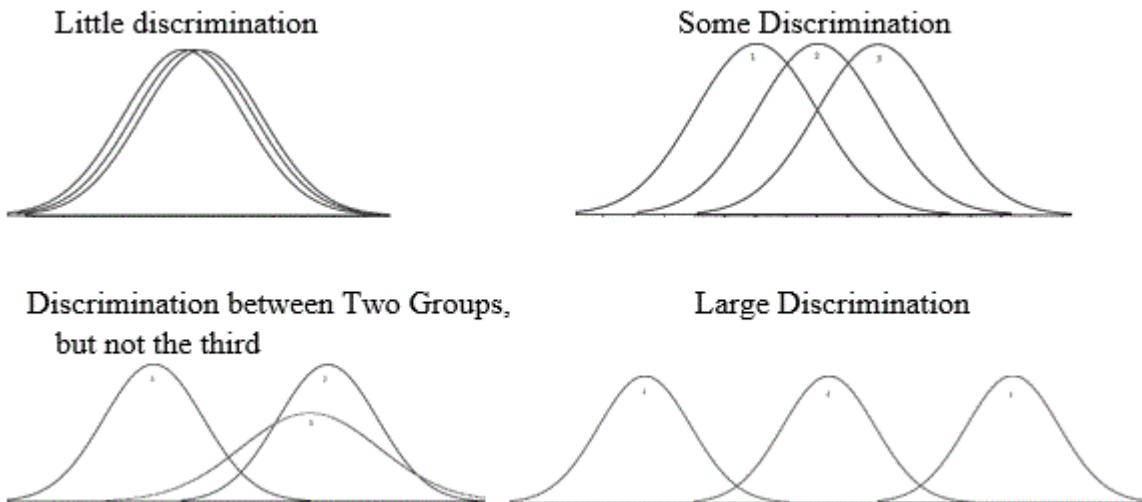
$$H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$$

$H_A:$ ***Not all μ values are equal***

- Note that the alternative hypothesis, ‘not all μ values are equal’, implies that some of them could be equal.
- The null hypothesis is equivalent to stating that the factor effects $\tau_1, \tau_2, \dots, \tau_k$ defined in the equation $Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$ are zero.

Setting up an ANOVA

The hypothesis test can be visualized as follows. Large discrimination in the means mean the factor levels have an impact on the outcome. If there is a little discrimination, it means that the factor levels don't have statistically significant impact.



For instance, suppose you are testing the effect of different diets on weight loss. If one diet results in an average weight loss of 10 kg and another results in an average weight loss of 1 kg, there's a large discrimination in the means. This would suggest that the type of diet (the factor) has a substantial effect on weight loss (the outcome).

Setting up an ANOVA

- We are interested in analyzing single factor effect with k levels, thus we will have k groups.

Let

k = Number of groups (or samples)

n_i = Number of observations in group i ($i = 1, 2, \dots, k$)

n = Total number of observations ($= \sum_{i=1}^k n_i$)

Y_{ij} = Observation j in group i

- μ_i = Mean of group i $= \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$

- μ = Overall mean $= \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} Y_{ij}$

Setting up an ANOVA

To arrive at the statistic , we calculate the following measures, which are variations within groups and between groups.

- **Sum of Squares of Total Variation (SST)**

Total variation is the sum of squared variation of **all values of response variable** (Y_{ij}) from the overall mean (μ) and is given by

$$SST = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \mu)^2$$

The degrees of freedom for SST is $(n - 1)$ since only the value of μ is estimated from n observations and thus only one degree of freedom is lost. Mean Square Total (MST) variation is given by

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

- Sum of Squares of Between (SSB) Group Variation

Sum of squares of between variation is the sum of squared variation between the group mean (μ_i) and the overall mean (μ) of the data and is given by

$$SSB = \sum_{i=1}^k n_i \times (\mu_i - \mu)^2$$

The degrees of freedom is $(k - 1)$. Since the overall mean μ is estimated from the data, one degree of freedom is lost. Mean square between variation (MSB) is given by

$$MSB = \frac{SSB}{k - 1}$$

- Sum of Squares of Within (SSW) Group Variation

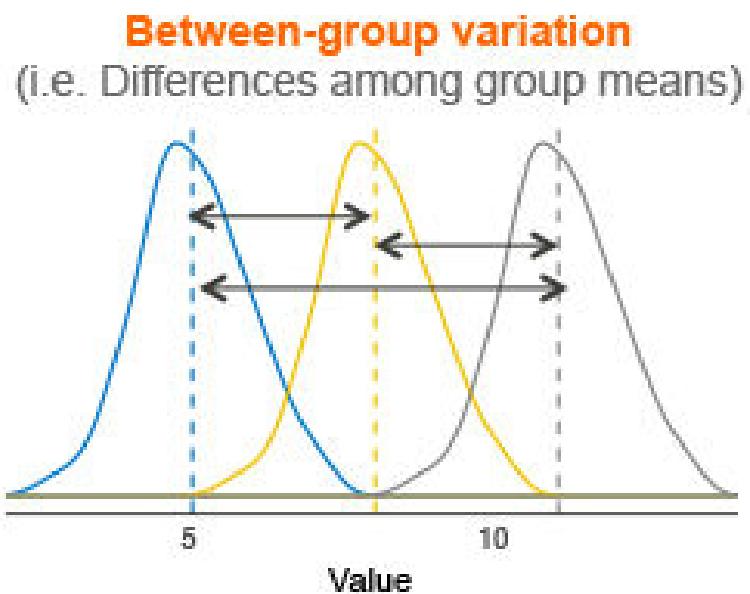
Sum of squares of within the group variation is the sum of squared variation of all observations (Y_{ij}) from that group mean (μ_i) and is given by

$$SSW = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \mu_i)^2$$

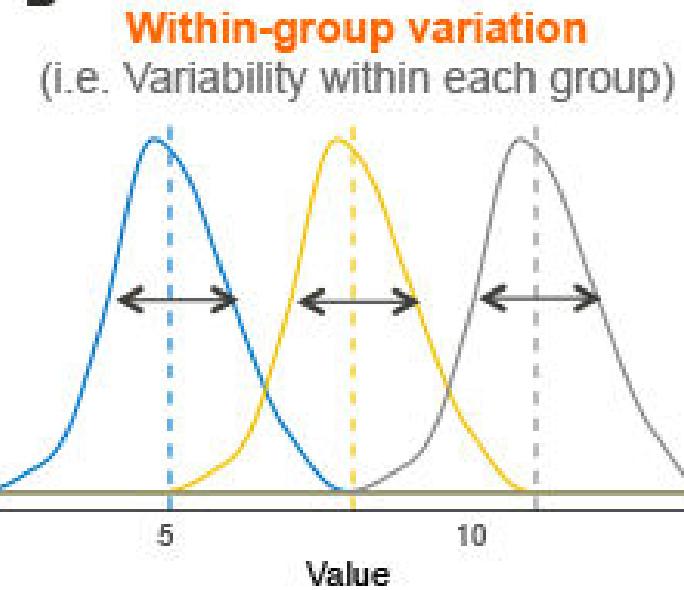
The degrees of freedom for SSW is $(n - k)$. Here k degrees of freedom are lost since we estimate k group means (μ_i). The mean square of variation within the group is

$$MSW = \frac{SSW}{n - k}$$

A



B



Setting up an ANOVA

We can prove algebraically

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \mu)^2 = \sum_{i=1}^k n_i \times (\mu_i - \mu)^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \mu_i)^2$$

That is,

$$SST = SSB + SSW$$

Cochran's Theorem

According to Cochran's theorem (Kutner *et al.*, 2013, page 70):

'If Y_1, Y_2, \dots, Y_n are drawn from a normal distribution with mean μ and standard deviation σ and sum of squares of total variation is decomposed into k sum of squares (SS_r) with degrees of freedom df_r , then the ratio (SS_r/σ^2) are independent χ^2 variables with df_r degrees of freedom if

$$\sum_{r=1}^k df_r = n - 1$$

Cochran's theorem ensures that the SSB and SSW are orthogonal, meaning they are independent of each other. This orthogonality simplifies the statistical analysis and facilitates the calculation of variance components.

Note that, in the equation $SST = SSB + SSW$, the SST is decomposed into two sums of squares (SSB and SSW) and thus, SSB/σ^2 and SSW/σ^2 are chi-square variables.

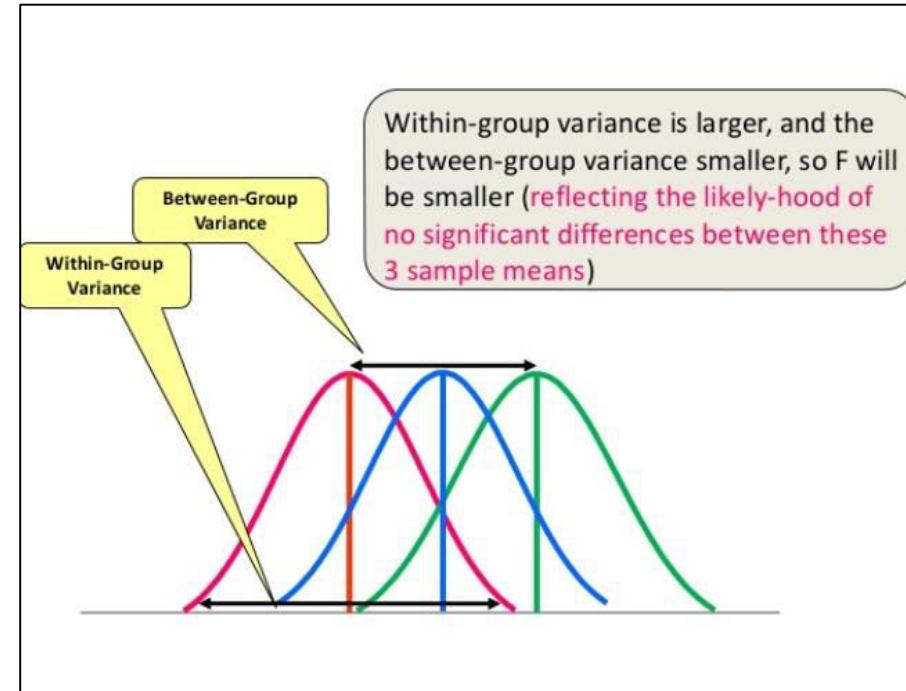
- If the null hypothesis is true, then there will be no difference in the mean values which will result in no difference between MSB and MSW .
- Alternatively, if the means are different, then MSB will be larger than MSW .
- That is the ratio MSB/MSW will be close to 1 if there is no difference between the mean values and will be larger than 1 if the means are different.

The F-test

- Following Cochran's theorem (Kirk, 1995) MSB/MSW is a ratio of two chi-square variate which is an F -distribution. Thus the statistic for testing the null hypothesis is

$$F = \frac{SSB/(k - 1)}{SSW/(n - k)} = \frac{MSB}{MSW}$$

- Note that the test statistic is a one-tailed test (right tailed) since we are interested in finding whether the variation between groups is greater than variation within the groups.
- It is important to note that rejecting the null hypothesis will not tell us exactly which means differ from each other , it will only indicate that there is a difference in at least one of the group means. We may have to conduct two-sample t-tests to find out which mean values are different.



Example (an Experimental Study)

Ms Rachael Khanna the brand manager of ENZO detergent powder at the ‘one stop’ retail was interested in understanding whether the price discounts has any impact on the sales quantity of ENZO. To test whether the price discounts had any impact, price discounts of 0% (no discount), 10% and 20% were given on randomly selected days. The quantity (in kilograms) of ENZO sold in a day under different discount levels is shown in below. Conduct a one-way ANOVA to check whether discount had any significant impact on the sales quantity at $\alpha = 0.05$.

No Discount (0% discount)									
39	32	25	25	37	28	26	26	40	29
37	34	28	36	38	38	34	31	39	36
34	25	33	26	33	26	26	27	32	40
10% Discount									
34	41	45	39	38	33	35	41	47	34
47	44	46	38	42	33	37	45	38	44
38	35	34	34	37	39	34	34	36	41
20% Discount									
42	43	44	46	41	52	43	42	50	41
41	47	55	55	47	48	41	42	45	48
40	50	52	43	47	55	49	46	55	42

Solution

- In this case, the number of groups $k = 3$; $n_1 = n_2 = n_3 = 30$; $\mu_1 = 32$, $\mu_2 = 38.77$, $\mu_3 = 46.4$; and $\mu = 39.05$.
- The sum of squares of between groups variation (SSB) is given by

$$\begin{aligned} SSB &= \sum_{i=1}^k n_i \times (\mu_i - \mu)^2 \\ &= 30 \times [(32 - 39.05)^2 + (38.77 - 39.05)^2 + (46.4 - 39.05)^2] = 3114.156 \end{aligned}$$

- Therefore

$$MSB = \frac{SSB}{k-1} = \frac{3114.156}{2} = 1557.078$$

- The sum of squares of within the group variation is given by

$$\begin{aligned}SSW &= \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \mu_i)^2 = \sum_{j=1}^{30} (Y_{1j} - 32)^2 + \sum_{j=1}^{30} (Y_{2j} - 38.77)^2 + \sum_{j=1}^{30} (Y_{3j} - 46.4)^2 \\&= 2056.567\end{aligned}$$

- Therefore

$$MSW = \frac{SSW}{n - k} = \frac{2056.567}{90 - 3} = 23.63$$

- The *F*-statistic value is

$$F_{2,87} = \frac{MSB}{MSW} = \frac{1557.078}{23.6387} = 65.86$$

- The critical F -value with degrees of freedom (2, 87) for $\alpha = 0.05$ is **3.101**
- The p -value for $F_{2,87} = 65.86$ is 3.82×10^{-18} (A small p -value (≤ 0.05) indicates strong evidence against the null hypothesis, so you reject the null hypothesis.)
- Since the calculated F -statistic is much higher than the critical F -value, we reject the null hypothesis and conclude that the mean sales quantity values under different discounts are different.

Anova: Single Factor						
SUMMARY						
Groups	Count	Sum	Average	Variance		
No Discount	30	960	32	27.17241		
10% Discount	30	1163	38.76667	20.46092		
20% Discount	30	1392	46.4	23.28276		
ANOVA						
Source of Variation	SS	df	MS	F	p-value	Fcrit
Between Groups	3114.15556	2	1557.078	65.86986	3.82E-18	3.101296
Within Groups	2056.56667	87	23.6387			
Total	5170.72222	89				

Excel output of ANOVA for this data

Test your understanding!

- What would happen if instead of using ANOVA to compare 7 groups , you performed multiple t-tests?
 - a) Making multiple comparisons with a t-test increases the probability of making a Type-1 error.
 - b) Sir Ronald Fischer would be turning over in his grave; He put all that work into developing ANOVA and you used multiple t-tests 😞
 - c) Nothing apart from making multiple comparisons with a t-test requires more computation than ANOVA
 - d) Nothing , both are the same.

Solution

- a) **Making multiple comparisons with a t-test increases the probability of making a Type-1 error.**
- What kind of a hypothesis test is used for one-way ANOVA?

Solution

Right-tailed test

Test your understanding!

- For an experiment with a single factor of k levels with n observations, the degrees of freedom for sum of squares of variation within the group is ?

- a) $n - 1$
- b) $k - 1$
- c) $n - k + 1$
- d) $n - k$

Solution

- d) $n - k$

Test your understanding!

- An investigator used ANOVA to compare 4 groups of students on numerical ability based on a class test. After analysis of raw scores, the following results were obtained. Calculate the value of the F-statistic

Source of Variation	DoF	Sum of squares
Between groups	3	625
Within Groups	36	2128

- a) 3.4
- b) 3.52
- c) 3.88
- d) 3.97

Solution:

3.52

Quick Glance-Points to remember

- Why ANOVA and what issue of the multiple T-test does it address?
- Mean model
- Factors effect model
- Setting up 1 way ANOVA:
 - Appropriate conditions where 1-way ANOVA is applicable
 - Understanding all the variables and subscripts used
 - Deriving SST,SSB and SSW(corresponding MST,MSB and MSW based on DoF)
 - Cochran's theorem
 - F-statistic for ANOVA
 - Finally, when to accept and reject the null hypothesis(based on calculated F value and critical F-value)

References

- Business Analytics by U. Dinesh Kumar – Wiley 2nd Edition, 2022
Chapter : 7.1 - 7.3.3
- <https://www.analyticsvidhya.com/blog/2018/01/anova-analysis-of-variance/>
- <https://www.analyticsvidhya.com/blog/2020/06/introduction-anova-statistics-data-science-covid-python/>
- <https://www.geeksforgeeks.org/one-way-anova/>



PES
UNIVERSITY

THANK YOU

Dr. Gowri Srinivasa

Professor, Department of Computer Science
and Engineering, PES University, Bengaluru

Email: gsrinivasa@pes.edu



DATA ANALYTICS

UE20CS312

UNIT-1

Lecture 9 : Analysis of Variance - 2

Gowri Srinivasa

Department of Computer Science and Engineering

Data Analytics

Unit 1

Lecture 9 : Analysis of Variance - 2

Slides excerpted from: U. Dinesh Kumar,
“Business Analytics”, Wiley, 2nd Edition 2022

Gowri Srinivasa

Department of Computer Science and Engineering

Slides collated by:

Slides collated by:

Nishanth M S, CSE 2023, PES University

nishanthmsathish.23@gmail.com

Harshitha Srikanth, VII CSE, PES University

harshithasrikanth13@gmail.com

With grateful thanks for contribution of slides to:

Dr. Mamatha H R, Professor at the Department of CSE, PESU

One-Way ANOVA : Example (Observational Study)

Share Raja Khan (SRK) is a top stockbroker and believes that the average annual stock return depends on the industrial sector. To validate his belief, SRK collected annual return of shares from three different industrial sectors – consumer goods, services, and industrial goods. The annual return of shares in 2015–2016 for different sectors is shown below. Conduct an ANOVA test at 5% significance level.

Annual return on 30 consumer goods stocks

6.32%	14.73%	11.95%	12.36%	10.28%	3.81%	10.15%	11.06%	6.29%	5.15%
8.44%	14.28%	8.89%	5.98%	6.96%	11.62%	5.22%	5.34%	5.93%	7.10%
10.91%	8.20%	10.19%	9.04%	8.61%	9.39%	2.63%	2.77%	4.76%	9.60%

Annual return on 30 services stocks

13.70%	3.58%	1.36%	17.41%	10.01%	10.88%	15.63%	-0.04%	10.32%	7.40%
11.48%	9.71%	11.19%	8.21%	1.64%	1.45%	10.12%	13.85%	-10.27%	5.26%
12.05%	4.47%	8.71%	5.59%	10.02%	7.65%	10.03%	7.87%	6.59%	13.60%

Annual return on 30 industrial goods stocks

6.74%	7.11%	5.69%	2.48%	5.42%	8.00%	2.55%	8.34%	4.99%	3.39%
8.73%	13.85%	5.29%	9.06%	2.84%	5.82%	7.66%	4.12%	9.10%	8.76%
10.77%	1.48%	4.71%	10.66%	0.44%	2.94%	6.55%	2.84%	3.90%	7.28%

Solution

- In this case, the number of cases $k = 3$; $n_1 = n_2 = n_3 = 30$; $\mu_1 = 0.082$, $\mu_2 = 0.079$, $\mu_3 = 0.0605$; and $\mu = 0.0743$
- The sum of squares of between (SSB) groups variation is given by

$$\begin{aligned} SSB &= \sum_{i=1}^k n_i \times (\mu_i - \mu)^2 \\ &= 30 \times [(0.082 - 0.0743)^2 + (0.079 - 0.0743)^2 + (0.0605 - 0.0743)^2] \\ &= 0.0087 \end{aligned}$$

- Therefore

$$MSB = \frac{SSB}{k-1} = \frac{0.0087}{2} = 0.0043$$

Solution

- The sum of squares of within the group variation is given by

$$\begin{aligned}
 SSW &= \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \mu_i)^2 \\
 &= \sum_{j=1}^{30} (Y_{1j} - 0.082)^2 + \sum_{j=1}^{30} (Y_{2j} - 0.079)^2 + \sum_{j=1}^{30} (Y_{3j} - 0.0605)^2 = 0.1463
 \end{aligned}$$

- Therefore

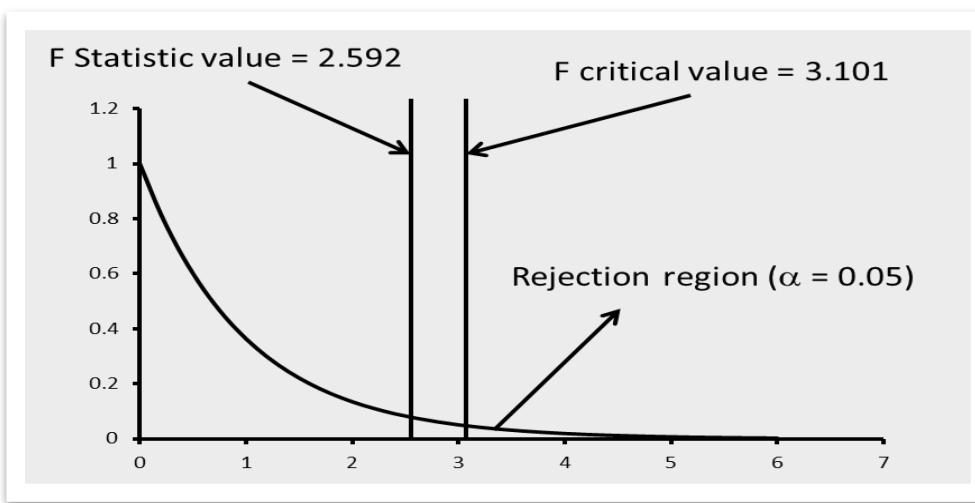
$$MSW = \frac{SSW}{n - k} = \frac{0.1463}{90 - 3} = 0.0016$$

- The F-statistic value is

$$F_{2,87} = \frac{MSB}{MSW} = \frac{0.0043}{0.0016} = 2.592$$

Solution

- The critical F -value with degrees of freedom (2, 87) for $\alpha = 0.05$ is 3.101
- The p -value for $F_{2,87} = 2.592$ is 0.0805 (A large p -value (> 0.05) suggests weak evidence against the null hypothesis, so you fail to reject the null hypothesis.)
- Since the calculated F -statistic is less than the critical F -value, we retain the null hypothesis and conclude that the average annual returns under industrial sectors consumer goods, services, and industrial goods are not different.



ANOVA: Single Factor						
SUMMARY						
Groups	Count	Sum	Average	Variance		
Consumer Goods	30	2.4796	0.082653	0.00101		
Services	30	2.3947	0.079823	0.003073		
Industrial Goods	30	1.8151	0.060503	0.000963		
ANOVA						
Source of Variation	SS	df	MS	F	p-value	Fcritical
Between Groups	0.008722	2	0.004361	2.59294	0.080572	3.101296
Within Groups	0.146317	87	0.001682			
Total	0.155039	89				

Excel output of ANOVA for this data

- The values of response variable may be influenced by several factors. For example, in addition to price discounts, location of the stores may also play an important role in the sales quantity.
- The discounts may not have much impact if the store is located near affluent community compared to stores located near non-affluent community.
- We would like to understand the impact of both factors (price discount and location) simultaneously on sales by trying to answer to the following questions:
 - Are there differences in the average sales quantity with different levels of price discounts?
 - Are there differences in the average sales quantity with respect to different locations?
 - Are there interactions between price discounts and location with respect to average sales quantity?

Setting up Two-Way ANOVA

The two-way ANOVA model can be expressed as

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha_i\beta_j + \varepsilon_{ijk}$$

Where,

Y_{ijk} = Value of the k^{th} observation ($k = 1, 2, \dots, K$) of the response variable at level i ($i = 1, 2, \dots, a$) of factor A and level j ($j = 1, 2, \dots, b$) of factor B.

μ = Overall mean value of the response variable Y_{ijk}

α_i = Level (effect) of factor A ($i = 1, 2, \dots, a$)

β_j = Level (effect) of factor B ($j = 1, 2, \dots, b$)

$\alpha_i\beta_j$ = Interaction of i^{th} level of factor A and j^{th} level of factor B

ε_{ijk} = Error associated with k^{th} of observation at level i of factor A and level j of factor B.

Setting up Two-Way ANOVA

The hypothesis tests associated with two-way ANOVA are as follows:

Test of Factor A Main Effects: Factor A with α levels

H_0 : $\alpha_i = 0$ for all i ($i = 1, 2, \dots, a$) There is no main effect of Factor A on the dependent variable.

H_A : Not all α_i are zero (At least one level of Factor A has a main effect that is different from zero.)

Test of Factor B Main Effects:

H_0 : $\beta_j = 0$ for all j ($j = 1, 2, \dots, b$)

H_A : Not all β_j are zero

Test of Interaction Effects:

H_0 : $\alpha_i\beta_j = 0$ for all i ($i = 1, 2, \dots, a$) and j ($j = 1, 2, \dots, b$)

H_A : Not all $\alpha_i\beta_j$ are zero

Setting up Two-Way ANOVA

The sum of squares in the case of two-way ANOVA with equal sample sizes is given by

$$SST = SSA + SSB + SSAB + SSW$$

Various components in the above equation are provided as follows :

- ***Sum of squared of total deviation (SST):***

$$SST = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^c (Y_{ijk} - \mu)^2$$

where c is the number of observations in each group and μ is the overall mean.

Setting up Two-Way ANOVA

- ***Sum of squares of deviation due to factor A (SSA):***

$$SSA = b \times c \times \sum_{i=1}^a (\mu_i - \mu)^2$$

where μ_i is the mean of all observations in level i of factor A and c is the number of observations in each group (assumed to be same for all groups).

- ***Sum of squares of deviation due to factor B (SSB):***

$$SSB = a \times c \times \sum_{j=1}^b (\mu_j - \mu)^2$$

Here μ_j is the mean of all observations in level j of factor B.

Setting up Two-Way ANOVA

- ***Sum of squares of deviation due to interaction of factors A and B (SSAB)***

$$SSAB = c \times \sum_{i=1}^a \sum_{j=1}^b (\mu_{ij} - \mu_i - \mu_j + \mu)^2$$

where μ_{ij} is the average of i^{th} level of factor A and j^{th} level of factor B.

- ***Sum of squares of deviation within a group (SSW):***

$$SSW = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^c (Y_{ijk} - \mu_{ij})^2$$

Setting up Two-Way ANOVA

Sum of squares of deviation for various effects and the corresponding *F*-statistic in a two-way ANOVA with equal sample size

Sum of Squared Variation	Degrees of Freedom	Mean Squared Variation	F-Statistics
SSA	a – 1	$MSA = SSA/(a - 1)$	$F = MSA/MSW$
SSB	b – 1	$MSB = SSB/(b - 1)$	$F = MSB/MSW$
SSAB	(a – 1)(b – 1)	$MSAB = SSAB/(a - 1)(b - 1)$	$F = MSAB/MSW$
SSW	ab(c – 1)	$MSW = SSW/ab(c - 1)$	

- Suppose a gardener wants to know if plant growth is influenced by sunlight exposure and watering frequency. She plants 40 seeds and lets them grow for one month under different conditions for sunlight exposure and watering frequency(Independent Variables).



		Sun light Exposure			
Watering Frequency		No light	Low	Medium	High
DAILY	DAILY	4.8	5	6.4	6.3
	DAILY	4.4	5.2	6.2	6.4
	DAILY	3.2	5.6	4.7	5.6
	DAILY	3.9	4.3	5.5	4.8
	DAILY	4.4	4.8	5.8	5.8
WEEKLY	WEEKLY	4.4	4.9	5.8	6
	WEEKLY	4.2	5.3	6.2	4.9
	WEEKLY	3.8	5.7	6.3	4.6
	WEEKLY	3.7	5.4	6.5	5.6
	WEEKLY	3.9	4.8	5.5	5.5

- The sum of squares in the case of two-way ANOVA with equal sample sizes is given by
- $SST = SSA + SSB + SSAB + SSW$
- A- First Factor (Watering Frequency)
- B- Second Factor (Sunlight Exposure)
- a- 2(Number of groups in Factor A- Watering frequency)
- b- 4(Number of groups in Factor B- Sunlight Exposure)
- c- 5(Number of observations in each group-ASSUMED TO BE SAME FOR ALL GROUPS)

Step 1: Calculation of All “mean”

$$\text{Mean of Daily} = (4.8 + 5 + 6.4 + 6.3 + \dots + 4.4 + 4.8 + 5.8 + 5.8) / 20 = 5.155$$

$$\text{Mean of Weekly} = (4.4 + 4.9 + 5.8 + 6 + \dots + 3.9 + 4.8 + 5.5 + 5.5) / 20 = 5.15$$

$$\text{Grand mean} = (4.8 + 5 + 6.4 + 6.3 + \dots + 3.9 + 4.8 + 5.5 + 5.5) / 40 = 5.1525$$

$$\text{MEAN(BOTH)} \quad 4.07 \quad 5.1 \quad 5.89 \quad 5.55$$

Step 2: Calculate Sum of Squares for First Factor A (Watering Frequency)

$$SSA = b \times c \times \sum_{i=1}^a (\mu_i - \mu)^2$$

where μ_i is the mean of all observations in level i of factor A and c is the number of observations in each group (assumed to be same for all groups).

$$SSA = 20(5.155-5.1525)^2 + 20(5.15-5.1525)^2 = .00025 \text{ ----- (Eq1)}$$

Step 3: Calculate Sum of Squares for Second Factor B (Sunlight Exposure)

$$SSB = a \times c \times \sum_{j=1}^b (\mu_j - \mu)^2$$

Here μ_j is the mean of all observations in level j of factor B.

$$SSB = 10(4.07-5.1525)^2 + 10(5.1-5.1525)^2 + 10(5.89-5.1525)^2 + 10(5.55-5.1525)^2 = 18.76475 \text{ ----- (Eq2)}$$

Step 4: Calculate Squares of Deviation(error)Within groups(SSW)



$$SSW = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^c (Y_{ijk} - \mu_{ij})^2$$

Next, we will calculate the sum of squares within by taking the sum of squared differences between each combination of factors and the individual plant heights.

- SS for daily watering and no sunlight: $(4.8-4.14)^2 + (4.4-4.14)^2 + (3.2-4.14)^2 + (3.9-4.14)^2 + (4.4-4.14)^2 = \mathbf{1.512}$
- SS for daily watering and low sunlight: **0.928**
- SS for daily watering and medium sunlight: **1.788**
- SS for daily watering and high sunlight: **1.648**
- SS for weekly watering and no sunlight: **0.34**
- SS for weekly watering and low sunlight: **0.548**
- SS for weekly watering and medium sunlight: **0.652**
- SS for weekly watering and high sunlight: **1.268**

Sums of squares within = $1.512 + .928 + 1.788 + 1.648 + .34 + .548 + .652 + 1.268 = \mathbf{8.684--}$
-----**(Eq3)**

Step 5: Calculate Total Sum of Squares

$$SST = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^c (Y_{ijk} - \mu)^2$$

where c is the number of observations in each group and μ is the overall mean.

Next, we can calculate the total sum of squares by taking the sum of the differences between each individual plant height and the grand mean:

Total Sum of Squares = $(4.8 - 5.1525)^2 + (5 - 5.1525)^2 + \dots + (5.5 - 5.1525)^2 = 28.45975$ -----
----- (Eq4)

Sum of Squared Variation	Degrees of Freedom	Mean Squared Variation	F-Statistics
SSA	$a - 1$	$MSA = SSA/(a - 1)$	$F = MSA/MSW$
SSB	$b - 1$	$MSB = SSB/(b - 1)$	$F = MSB/MSW$
SSAB	$(a - 1)(b - 1)$	$MSAB = SSAB/(a - 1)(b - 1)$	$F = MSAB/MSW$
SSW	$ab(c - 1)$	$MSW = SSW/ab(c - 1)$	

Let us fill the values in our 2-way ANOVA table: alpha= 0.05

Sum of Squared Variation	Degrees of Freedom	Mean Squared Variation	F-Statistics	F-Critical	p-value
SSA(watering frequency) =0.00025	2 - 1=1	MSA = 0.00025/1=0.00025	F = MSA/MSW=0.000921	= (DFssa,DFwithin) =4.14910	0.975
SSB(Sunlight exposure)=18.76475	4 - 1=3	MSB = 18.76475/3=6.254917	F = MSB/MSW=23.04898	F(3,32)=0	<.000
SSAB (Interaction)=1.01075	(2 - 1)(4 - 1)=3	MSAB = 1.01075/3=0.336917	F = MSAB/MSW=1.241517	F(3,32)=0	0.311
SSW(within)=8.684	2*4(5 - 1)=32	MSW = 8.684/32=0.271375			
SST		28.45975			

Step 8: Interpret the results

We can observe the following from the ANOVA table:

- The p-value for the interaction between watering frequency and sunlight exposure was **0.311**. This is not statistically significant at $\alpha = 0.05$. Hence
- The p-value for watering frequency was **0.975**. This is not statistically significant at $\alpha = 0.05$.
- The p-value for sunlight exposure was **< 0.000**. This is statistically significant at $\alpha = 0.05$.

If $p\text{-value} < 0.05$ – “significant” - reject the null hypothesis

$p\text{-value} > 0.05$ – not statistically significant - retain the null hypothesis

Two-Way ANOVA : Example

The table next slide shows the sales quantity of detergents at different discount values and different locations collected over 20 days. Conduct a two-way ANOVA at $\alpha = 0.05$ to test the effects of discounts and location on the sales.

Location 1			Location 2		
Discount			Discount		
0%	10%	20%	0%	10%	20%
20	28	32	20	19	20
16	23	29	21	27	31
24	25	28	23	23	35
20	31	27	19	30	25
19	25	30	25	25	31
10	24	26	22	21	31
24	28	37	25	33	31
16	23	33	21	26	23
25	26	27	26	22	22
16	25	31	22	28	32
18	22	37	25	24	22
20	24	28	23	23	29
17	26	25	23	26	25
26	28	23	24	16	34
16	21	26	20	30	30
21	27	33	23	22	25
24	25	28	18	16	39

DATA ANALYTICS



Solution

ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Sample (Location)	7.008333	1	7.008333	0.443898	0.506593	3.92433
Columns (Discount)	1240.317	2	620.1583	39.27997	1.06E-13	3.075853
Interaction	84.81667	2	42.40833	2.686085	0.07246	3.075853
Within	1799.85	114	15.78816			
Total	3131.992	119				

Solution

- In the table , the sample stands for the row factor (which in this case is location), column stands for the column factor (discount in this case), and interaction stands for interaction effect (location \times discount).
- The *p*-value for locations (data in rows) is 0.5065, thus it is not statistically significant (we retain the null hypothesis that the locations have no statistical influence on sales), **whereas for discount rates (data in column) the *p*-value is 1.06×10^{-13} , so we reject the null hypothesis (that is discount rate has influence on sales)**.
- The *p*-value for the interaction effect is 0.0724 and is not significant. That is only the factor discount is statistically significant at $\alpha = 0.05$.

- Analysis of Variance (ANOVA) is a hypothesis testing procedure used for comparing means from several groups simultaneously.
- In one-way ANOVA, we test whether the mean values of an outcome variable for different levels of factor are different. Using multiple two sample t-test to simultaneously test group means will result in incorrect estimation of Type- I error and ANOVA overcomes this problem.
- ANOVA plays an important role in multiple linear regression model diagnostics. The overall significance of the model is tested using ANOVA.
- In a two-way ANOVA we check the impact of more than one factor simultaneously on several groups.

Test your understanding!

- A two-way ANOVA means that the experimental design includes
 - a) Two dependent variables
 - b) Two independent variables
 - c) Two types of variance
 - d) All of these

Solution

- b) Two independent variables.**

Test your understanding!

- In a two-way ANOVA, the interaction effect is the
 - a) Effect of changing the levels of one factor on the dependent scores
 - b) Effect of changing the levels of one factor on the dependent scores, ignoring all other factors in the study
 - c) Extent to which the influence one factor has on scores depends on the level of the other factor
 - d) Effect on the independent variables of changing the levels of a factor

Solution

- c) Extent to which the influence one factor has on scores depends on the level of the other factor**

Quick Glance-Points to remember

- Why 2-way ANOVA(2 factors considered)
- Setting up 2-way ANOVA:
 - Understanding all the variables and subscripts used
 - Test of Factor A(Main effect)
 - Test of Factor B(Main effect)
 - Interaction effect
 - SST,SSA,SSB,SSAB,SSW and corresponding DoF's
 - F-statistics
 - Finally, when to accept and reject the null hypothesis(based on calculated F value and critical F-value and p value)

References

- Business Analytics by U. Dinesh Kumar – Wiley 2nd Edition, 2022
Chapter : 7.4
- <https://www.analyticsvidhya.com/blog/2018/01/anova-analysis-of-variance/>
- <https://www.analyticsvidhya.com/blog/2020/06/introduction-anova-statistics-data-science-covid-python/>



THANK YOU

Dr. Gowri Srinivasa

Professor, Department of Computer Science
and Engineering, PES University, Bengaluru

Email: gsrinivasa@pes.edu



PES
UNIVERSITY

DATA ANALYTICS

UE21CS342AA2

UNIT-1

Lecture 11 : Correlation Analysis - 1

Gowri Srinivasa

Department of Computer Science and Engineering

Data Analytics

Unit 1

Lecture 11 : Correlation Analysis - 1

Slides excerpted from: U. Dinesh Kumar,
“Business Analytics”, Wiley, 2nd Edition 2022

Gowri Srinivasa

Department of Computer Science and Engineering

Slides collated by:

Nishanth M S, CSE 2023, PES University
nishanthmsathish.23@gmail.com

Harshitha Srikanth, VII CSE, PES University
harshithasrikanth13@gmail.com

With grateful thanks for contribution of slides to:
Dr. Mamatha H R, Professor at the Department of CSE, PESU

Need for Correlation

- Organizations collect data on **several variables** and generate several more through feature engineering which may bring its count to thousands.
- One of the most challenging tasks in predictive analytics is **identifying features** that may be associated with the response or the **outcome variable** of interest.
- It is important **to reduce the number of features** to improve the time taken for analysis as well as ensure a model isn't destabilised due to excess features.

Introduction to Correlation

- Correlation is a statistical measure of an associative relationship that exists between two random variables.
- It is a *measure of strength and direction* of a relationship that exists between two random variables and is measured using a correlation coefficient r .
- It is not necessarily a causal relationship.
- Correlation is important in analytics as it helps to identify variables that may be used in the model building and also useful for identifying issues such as multi-collinearity that can destabilise regression-based models.
- It is also useful for finding proxy variables in analytics model building.

Multi-collinearity occurs when two or more predictor variables in a regression model are highly correlated. In the presence of multi-collinearity, it becomes difficult to determine the individual impact of each predictor variable on the dependent variable.

Pearson Correlation Coefficient

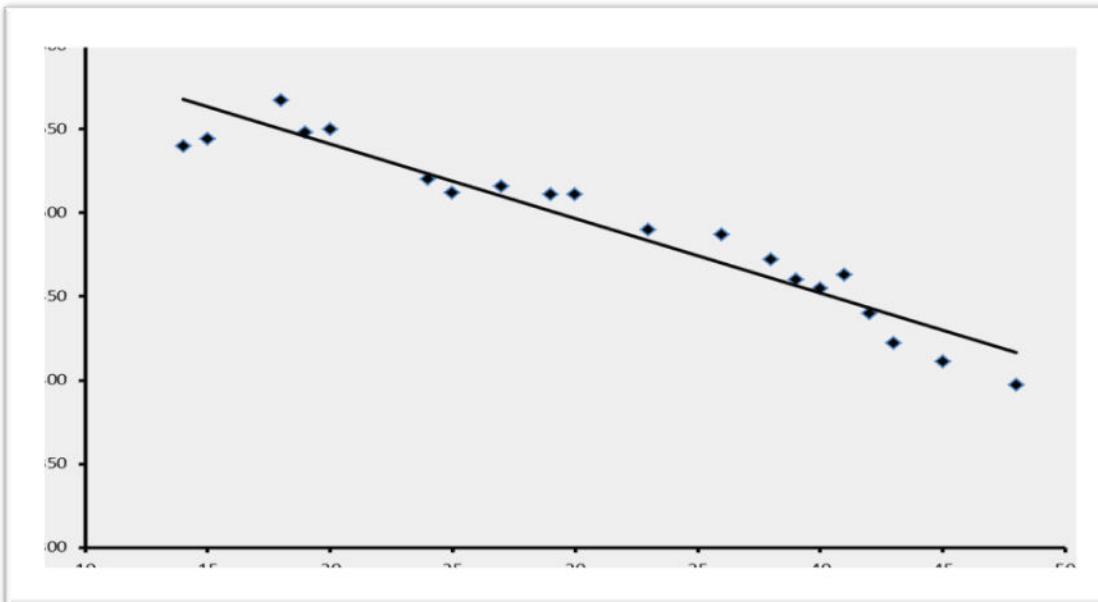
- Pearson product moment correlation (in short Pearson correlation) is used for measuring the **strength** and **direction** of the **linear** relationship between two **continuous random variables**.

Consider the following data

Age	14	15	18	19	20	24	25	27	29	30
Call Duration	540	544	567	548	550	520	512	516	511	511
Age	33	36	38	39	40	41	42	43	45	48
Call Duration	490	487	472	460	455	463	440	422	411	397

Pearson Correlation Coefficient

In the graph , we can see that the average call duration (Y) decreases as the age of the customer (X) increases. We can measure the strength of the linear association relationship using a numerical measure called correlation coefficient.



In other words, as one variable increases or decreases by a certain amount, the other variable also changes by a consistent multiple of that amount. This change follows a straight-line pattern when plotted on a graph.

$$Y=a+bX$$

Association relationship between age and average call duration

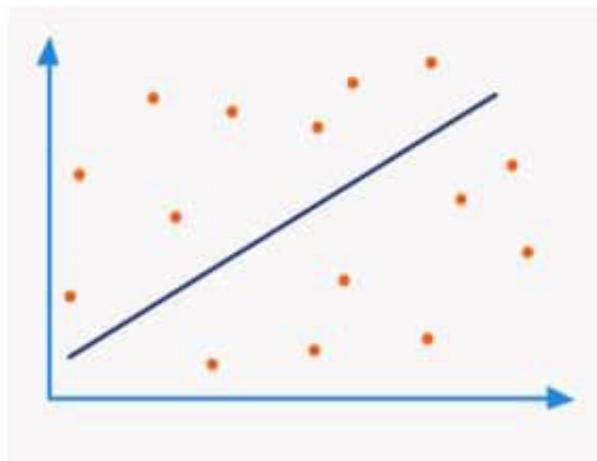
1.
Large positive
correlation



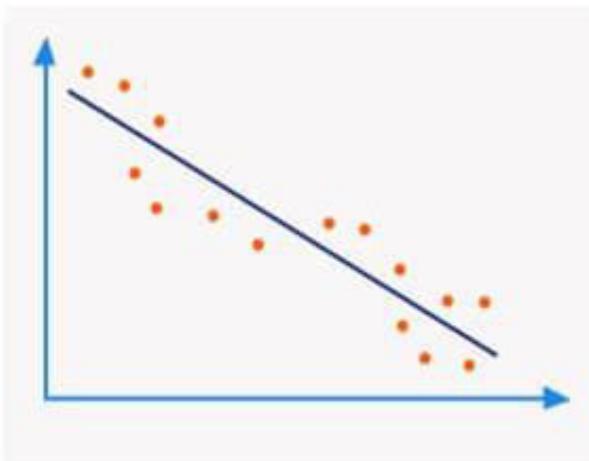
2.
Medium positive
correlation



4.
Weak / no
correlation



3.
Small negative
correlation



Calculating Pearson Correlation Coefficient

- To calculate Pearson correlation coefficient , the variables must belong to *ratio or interval* scale. The range of the variables can be different, thus we need to standardize the variables which then can be used to find the correlation.
- Let X_i be different values of the variable X and Y_i be different values of Y . Then the standardized values of X and Y are given by

$$Z_X = \left(\frac{X_i - \bar{X}}{\sigma_X} \right) \quad Z_Y = \left(\frac{Y_i - \bar{Y}}{\sigma_Y} \right)$$

Where \bar{X} and \bar{Y} are mean values of the random variables X and Y . σ_X and σ_Y are the corresponding standard deviations.

Equal Intervals: Interval and ratio scales have equal intervals between values. This property is essential for interpreting the correlation coefficient. The coefficient measures how much the variables change together on average for a unit change in one variable. For example, if one variable goes from 1 to 2 and another variable goes from 5 to 6, the change is the same (1 unit) for both variables.

Calculating Pearson Correlation Coefficient

The Pearson's correlation coefficient is given by

$$r = \frac{\sum_{i=1}^n Z_X Z_Y}{n} = \frac{\sum_{i=1}^n (X_i - \bar{X}) \times (Y_i - \bar{Y})}{n \sigma_X \sigma_Y}$$

Where n is the number of records in the population. In case of working with the sample , to account for the degrees of freedom , the following formula is recommended

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X}) \times (Y_i - \bar{Y})}{(n - 1) S_X S_Y}$$

For large number of samples , the correlation coefficients calculated by both these equations converge.

$$r = \frac{\sum_{i=1}^n Z_X Z_Y}{n} = \frac{\sum_{i=1}^n (X_i - \bar{X}) \times (Y_i - \bar{Y})}{n \sigma_X \sigma_Y}$$

- Whenever the value of X_i is greater than mean and if the corresponding value of Y_i is also greater than mean, then the numerator in equation will be positive.
- Whenever the value of X_i is lesser than mean and if the corresponding value of Y_i is also lesser than mean, then the numerator in equation will be positive.
- Whenever the value of X_i is lesser than mean (or greater than mean) and the corresponding value of Y_i is greater than mean (or lesser than mean), then the numerator in equation will be negative.
- It is possible that we may have combinations of three cases listed above in a data set. Thus the numerator in the equation is likely to be positive, negative, or zero.

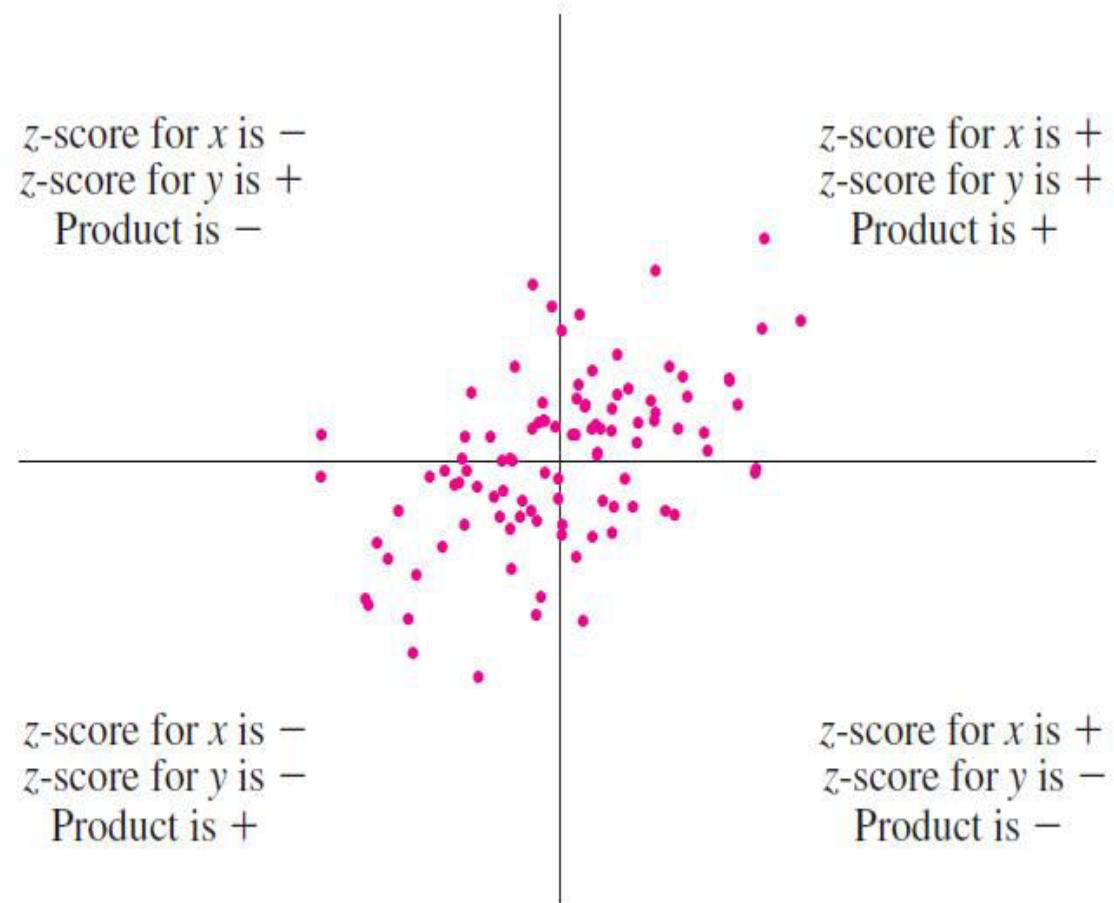


FIGURE 7.5 How the correlation coefficient works.

Equivalent Equations

- $r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \times \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$
- $r = \frac{n \sum_{i=1}^n X_i Y_i - \sum_{i=1}^n X_i \sum_{i=1}^n Y_i}{\sqrt{n \sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2} \times \sqrt{n \sum_{i=1}^n Y_i^2 - (\sum_{i=1}^n Y_i)^2}}$
- $r = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$

Where $\text{Cov}(X,Y)$ is the covariance between the random variables X and Y

and is given by $\text{Cov}(X,Y) = E((X_i - \bar{X})(Y_i - \bar{Y}))$

Properties of Pearson Correlation Coefficient

- The value of correlation coefficient lies between -1 and $+1$. High absolute value of r , $|r|$, indicates strong relationship between the two variables.
- Positive value of r indicates positive correlation (as value of X increases, the value of Y also increases) and negative value of r indicates negative correlation (as the value of X increases, the value of Y decreases).
- The sign of correlation coefficient is same as the sign of covariance between the two random variables.
- The Correlation Coefficient Is Unitless

Summary

The correlation coefficient remains unchanged under each of the following operations:

- Multiplying each value of a variable by a positive constant.
- Adding a constant to each value of a variable.
- Interchanging the values of x and y .

Properties of Pearson Correlation Coefficient

- Assume that the value of Pearson correlation coefficient between X and Y is r . Let Z_1 and Z_2 be the linear combinations of X and Y ($Z_1 = A + BX$ and $Z_2 = C + DY$). Then the correlation coefficient between Z_1 and Z_2 will be r when the signs of B and D are same (both are positive or negative) and $-r$ when the signs of B and D are opposite.
- Mathematically, square of correlation coefficient is equal to the co-efficient of determination (R^2) of the linear regression model, that is $r^2 = R^2$.
- Pearson correlation coefficient value may be zero even when there is a strong non-linear relationship between variables X and Y . Thus low correlation coefficient value cannot be taken as an evidence of no relationship.

Properties of Pearson Correlation Coefficient

- Assume that the value of Pearson correlation coefficient between X and Y is r . Let Z_1 and Z_2 be the linear combinations of X and Y ($Z_1 = A + BX$ and $Z_2 = C + DY$). Then the correlation coefficient between Z_1 and Z_2 will be r when the signs of B and D are same (both are positive or negative) and $-r$ when the signs of B and D are opposite.
- This property emphasizes that the correlation coefficient remains invariant under linear transformations, but its sign can change based on the directions of those transformations.

Example:

The correlation coefficient between two variables X and Y is found to be 0.6. All the observations on X and Y are transformed using the transformations $U = 2 - 3X$ and $V = 4Y + 1$. The correlation coefficient between the transformed variables U and V will be

Example

The average share prices of two companies over the past 12 months are shown below. Calculate the Pearson Correlation coefficient.

X	Y
274.58	219.50
287.96	242.92
290.35	245.90
320.07	256.80
317.40	240.60
319.53	245.23
301.52	232.09
271.75	222.65
323.65	231.74
259.80	214.43
263.02	201.86
286.03	204.23

The average values are :

$$\bar{X} = 292.9717 \quad \bar{Y} = 229.8292$$

Correlation coefficient is calculated using the following equation:

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \times \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

DATA ANALYTICS

Example



X_i	Y_i					
274.58	219.50	-18.39	-10.33	189.97	338.25	106.6917
287.96	242.92	-5.01	13.09	-65.61	25.12	171.3699
290.35	245.90	-2.62	16.07	-42.13	6.87	258.2717
320.07	256.80	27.10	26.97	730.86	734.32	727.4259
317.40	240.60	24.43	10.77	263.11	596.74	116.0109
319.53	245.23	26.56	15.40	409.02	705.35	237.1857
301.52	232.09	8.55	2.26	19.33	73.07	5.111367
271.75	222.65	-21.22	-7.18	152.35	450.36	51.54043
323.65	231.74	30.68	1.91	58.62	941.16	3.651284
259.80	214.43	-33.17	-15.40	510.82	1100.36	237.1343
263.02	201.86	-29.95	-27.97	837.72	897.10	782.2743
266.66	201.86	-29.95	-27.97	837.72	897.10	782.2743

Example

From the table , we have

- $\sum_{i=1}^{12}(X_i - \bar{X})(Y_i - \bar{Y}) = 3241.77$

- $\sum_{i=1}^{12}(X_i - \bar{X})^2 = 5916.89$

- $\sum_{i=1}^{12}(Y_i - \bar{Y})^2 = 3351.98$

Correlation coefficient $r = \frac{3241.77}{\sqrt{5916.89} \times \sqrt{3351.98}} = 0.7279$

Hypothesis test for Correlation Coefficient

For any two sets of data , the Pearson correlation coefficient is most likely to give a value other than zero. Many thumb rules exist to group the correlation value as no correlation, low correlation, medium correlation, and high correlation (Monroe and Stuit, 1933).

Let ρ be the population correlation coefficient. The null and alternative hypotheses are given by

$H_0:$	$\rho = 0$ (there is no correlation between two random variables)
$H_A:$	$\rho \neq 0$ (there is a correlation between two random variables)

Hypothesis test for Correlation Coefficient

- The sampling distribution of correlation coefficient r follows an approximate t -distribution with $(n - 2)$ degrees of freedom (dof) where n is the number of cases in the sample for calculating the correlation coefficient.
- Two degrees of freedom are lost since we estimate two mean values from the data. The mean of the sampling distribution is ρ and the corresponding

standard deviation is (Ezekiel, 1941) $\sqrt{\frac{1-r^2}{n-2}}$

- The t -statistic for null hypothesis is given by $t_{\alpha/2,n-2} = \frac{r-\rho}{\sqrt{\frac{1-r^2}{n-2}}}$
- When the null hypothesis is $\rho = 0$, the test statistic in above equation

becomes $t_{\alpha/2,n-2} = r \sqrt{\frac{n-2}{1-r^2}}$

Hypothesis test for Correlation Coefficient - Example

The average share prices of two companies over the past 12 months are shown in Table. conduct the following two hypothesis tests at $\alpha = 0.05$:

- (a) The correlation between share prices of two companies is zero.
- (b) The correlation between share prices of two companies is at least 0.5.

X	Y
274.58	219.50
287.96	242.92
290.35	245.90
320.07	256.80
317.40	240.60
319.53	245.23
301.52	232.09
271.75	222.65
323.65	231.74
259.80	214.43
263.02	201.86
286.03	204.23

•a) The null and alternative hypotheses are:

$$H_0: \rho = 0$$

$$H_A: \rho \neq 0$$

The corresponding t -statistic is

$$t = r \sqrt{\frac{n - 2}{1 - r^2}} = 0.7279 \sqrt{\frac{12 - 2}{1 - 0.7279^2}} = 3.3569$$

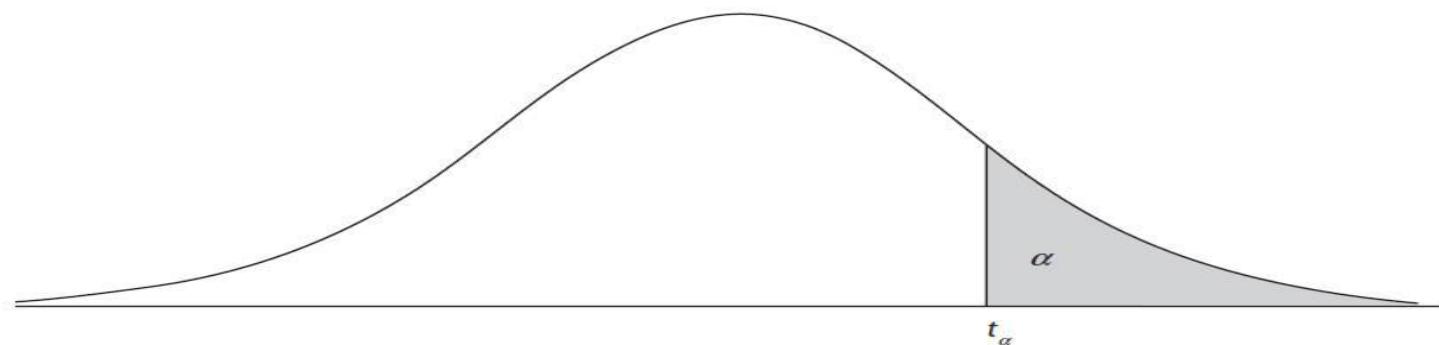
Note that this is a **two-tailed test** and the critical t -value at $\alpha = 0.05$ and $df = 10$ is 2.2281. (here α will be $\alpha/2=0.025$)

Since the calculated t -statistic is higher than the critical t -value, we reject the null hypothesis and conclude that **there is a significant correlation between share prices of two companies.**

The corresponding p -value is 0.0072.

Note that this is a two-tailed test and the critical t-value at alpha = 0.05 and df = 10(n-2) is 2.2281

t-Distribution Critical Values



Area to the right of: α

Degrees of freedom	$t_{0.100}$	$t_{0.050}$	$t_{0.025}$	$t_{0.01}$	$t_{0.005}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169

•b) The null and alternative hypotheses are given by

$$H_0: \rho \leq 0.5$$

$$H_A: \rho > 0.5$$

The corresponding t -statistic is

$$t = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}} = \frac{0.7279 - 0.5}{\sqrt{\frac{0.2168}{n - 2}}} = 1.05$$

This is a right-tailed test and the corresponding t -critical value is 1.8124.

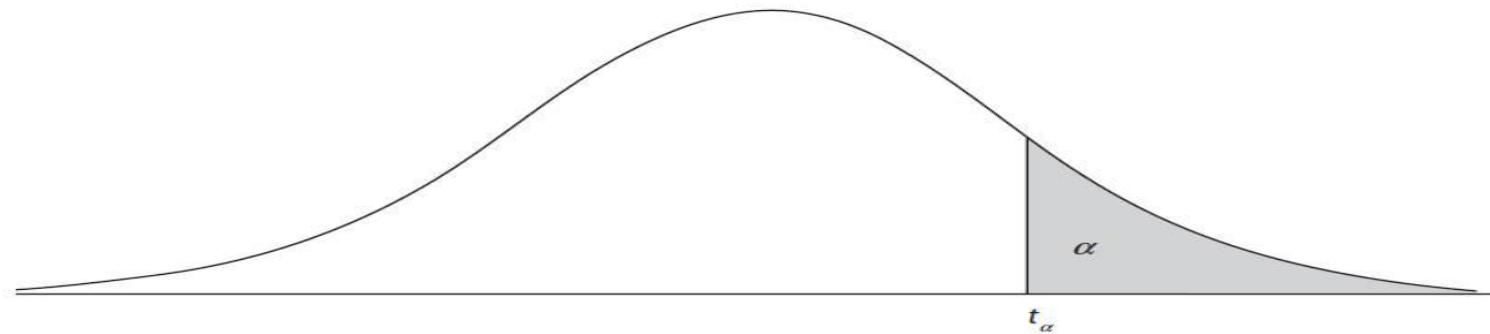
The calculated t -value is less than the critical value of t , and thus we retain the null hypothesis and conclude that the correlation between share prices of two companies is less than 0.5.

The corresponding p -value is 0.1592.

Right-tailed test and the corresponding t-critical value is 1.8124.



t-Distribution Critical Values



Area to the right of : α

Degrees of freedom	$t_{0.100}$	$t_{0.050}$	$t_{0.025}$	$t_{0.01}$	$t_{0.005}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169



P value(right tailed)

TABLE D *t* distribution critical values

df	Upper tail probability <i>p</i>											
	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073

Test your understanding!

- Pearson correlation is applicable to what type of attribute(s)?

Solution

Attributes from ratio scale or interval scale

- What is the range of Pearson correlation coefficient?

Solution

[-1,1]

- Pearson correlation between 2 variables X and Y is 0.85. What is the Pearson correlation between $3X-100$ and $400-10Y$?

Solution

-0.85

References

- Business Analytics by U. Dinesh Kumar – Wiley 2nd Edition, 2022
Chapters : 8.1 – 8.2.4



THANK YOU

Dr. Gowri Srinivasa

Professor, Department of Computer Science
and Engineering, PES University, Bengaluru

Email: gsrinivasa@pes.edu



PES
UNIVERSITY

DATA ANALYTICS

UE21CS342AA2

UNIT-1

Lecture 12 : Correlation Analysis - 2

Gowri Srinivasa

Department of Computer Science and Engineering

Data Analytics

Unit 1

Lecture 12 : Correlation Analysis - 2

Slides excerpted from: U. Dinesh Kumar,
“Business Analytics”, Wiley, 2nd Edition 2022

Gowri Srinivasa

Department of Computer Science and Engineering

Slides collated by:

Nishanth M S, CSE 2023, PES University
nishanthmsathish.23@gmail.com

Harshitha Srikanth, VII CSE, PES University
harshithasrikanth13@gmail.com

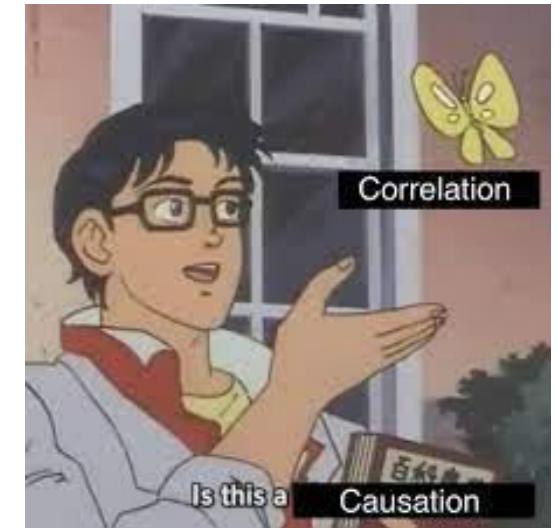
With grateful thanks for contribution of slides to:
Dr. Mamatha H R, Professor at the Department of CSE, PESU

Spurious Correlation

- One of the major problem with correlation is the possibility of spurious correlation between two random variables which in many cases is caused due to some other latent variable (hidden variable) that influences both variables for which the correlation is calculated.

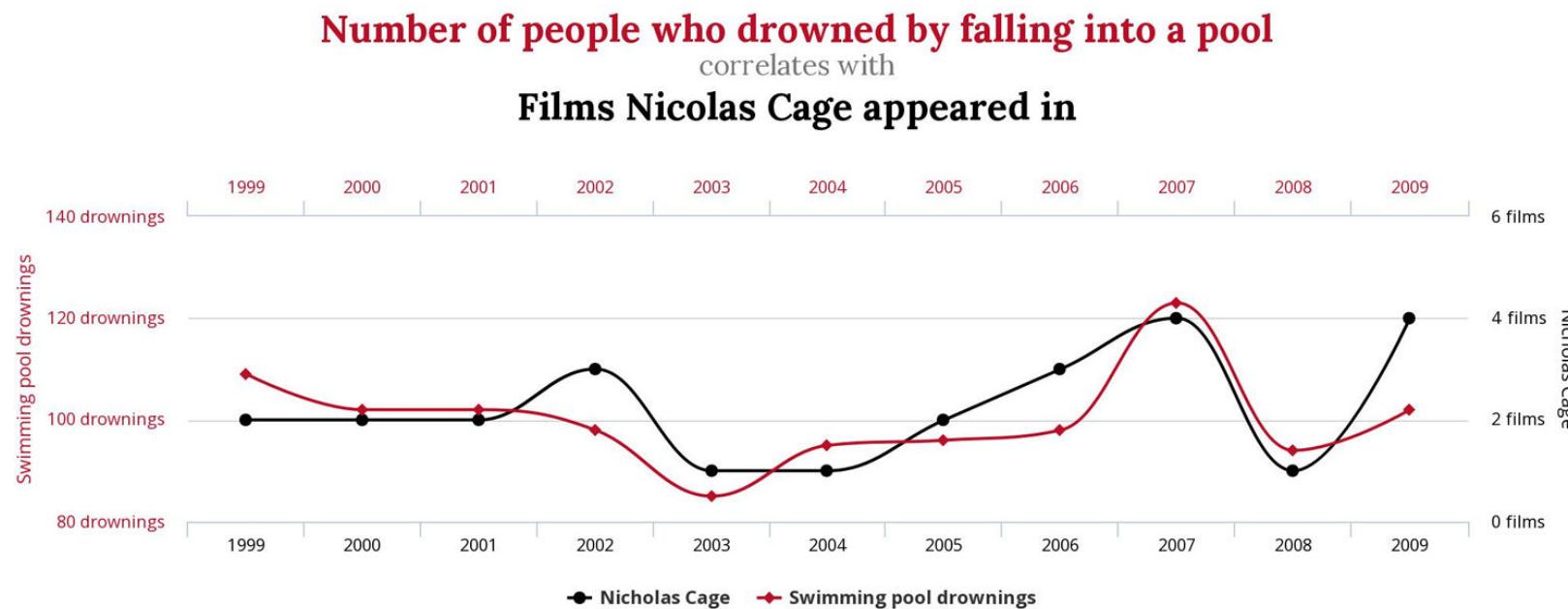
Following are few examples of spurious correlation between two random variables:

- **Crime rate versus ice cream sale:** It has been reported that the sale of ice cream and crime rates are positively correlated (Levitt and Dubner, 2009). Obviously, ice cream is not driving the crime rate. In this case the hidden variable is the temperature – Ice cream sales increase during the summer and since summer is the time of vacation , many homes are left empty , susceptible for theft thus driving the crime rates up.



Spurious Correlation

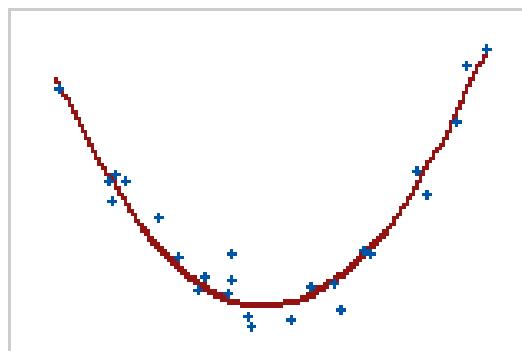
- Doctors and deaths:** Number of doctors is positively correlated with number of deaths in villages, that is, as the number of doctors increases, the deaths also increase. We can be sure that doctors are not causing the deaths to increase (Young, 2001). In this case , the need for doctors is more in villages with poor public healthcare.



Should Niclos Cage end his career to save lives?

Pearson Correlation with a non-linear relationship

- Pearson correlation finds a linear relationship between 2 random variables. If the resultant correlation is less between two variables, one cannot conclude the variables aren't correlated.
- For example, the correlation for a set of data turned out to be only 0.224. But on plotting a scatter plot , it was observed that the curve could be better modelled using a non-linear function like a quadratic or a cubic function. Thus it is always advisable to validate your results by visualizing the data.



A nonlinear relationship

The Correlation Coefficient can be Misleading when Outliers are Present

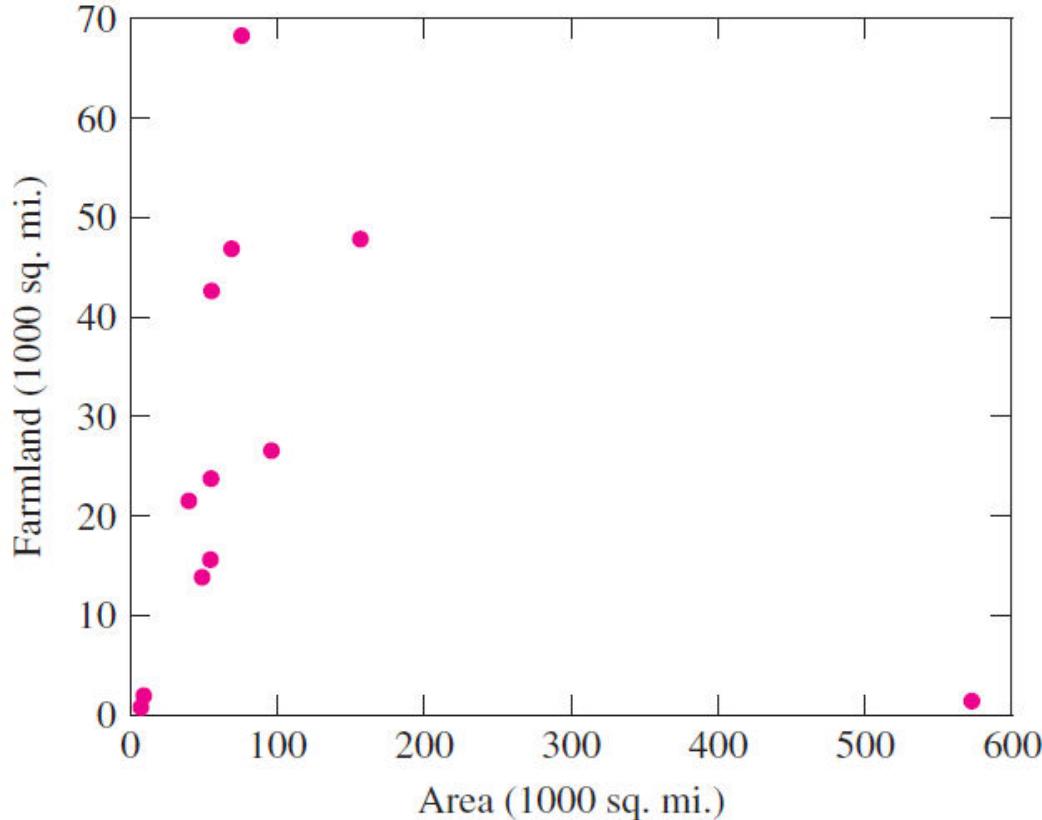


FIGURE 7.8 The correlation is -0.12 . Because of the outlier, the correlation coefficient is misleading.

Spearman Rank Correlation

- Pearson correlation is appropriate when the random variables involved are both from either ratio scale or interval scale.
- When both random variables are of **ordinal scale**, we use **Spearman rank correlation (also known as Spearman's rho denoted by ρ_s)**.
- The Spearman rank correlation, r_s , estimated from a sample is given by (Yule and Kendall 1937, Woodbury, 1940)

$$r_s = 1 - \frac{6 \sum_{i=1}^n D_i^2}{n(n^2 - 1)}$$

where D_i = difference in the rank of case i under variables X and Y (that is $X_i - Y_i$).

- The sampling distribution of Spearman correlation r_s also follows an approximate t -distribution with mean ρ_s and standard deviation $\sqrt{\frac{1-r_s^2}{n-2}}$ with $n - 2$ degrees of freedom

Example

Ranking of 12 countries under corruption and Gini Index (wealth discrimination) are shown in the table. Calculate the Spearman correlation and test the hypothesis that the correlation is at least 0.2 at $\alpha = 0.02$.

Countries	1	2	3	4	5	6	7	8	9	10	11	12
Corruption	1	4	12	2	5	8	11	7	10	3	6	9
Gini Index	2	3	9	5	4	6	10	7	8	1	11	12

The calculations are as follows

Country	Corruption Rank (X_i)	Gini Index (Y_i)	$D = X_i - Y_i$	D^2
1	1	2	-1	1
2	4	3	1	1
3	12	9	3	9
4	2	5	-3	9
5	5	4	1	1
6	8	6	2	4
7	11	10	1	1
8	7	7	0	0
9	10	8	2	4
10	3	1	2	4
11	6	11	-5	25
12	9	12	-3	9

- The Spearman rank correlation is given by

$$r_s = 1 - \frac{6 \sum_{i=1}^n D_i^2}{n(n^2 - 1)} = 1 - \frac{6 \times 68}{12(12^2 - 1)} = 0.7622$$

- The null and alternative hypotheses are

$$H_0: \rho_s \leq 0.2$$

$$H_A: \rho_s > 0.2$$

The corresponding t -statistic is

$$t = \frac{r_s - \rho_s}{\sqrt{\frac{1 - r_s^2}{n - 2}}} = \frac{0.7622 - 0.2}{\sqrt{\frac{1 - 0.7622^2}{12 - 2}}} = 2.74$$

- The one-tailed t -critical value for $\alpha = 0.02$ and $dof = 10$ is 2.35. Since the calculated t -statistic value is more than the t -critical value, we reject the null hypothesis and conclude that Spearman rank correlation between two countries is at least 0.2.

- **Point Bi-Serial correlation** is used when we are interested in finding correlation between a **continuous random variable** and a **dichotomous (binary) random variable**.
- Assume that the random variable X is a continuous random variable and Y is a dichotomous random variable. Then the following steps are used for calculating the correlation between these two variables:
 1. Group the data into two sets based on the value of the dichotomous variable Y . That is, assume that the value of Y is either 0 or 1. Then we group the data into two subsets such that in one group the value of Y is 0 and in another group the value of Y is 1.
 2. Calculate the mean values of two groups: Let \bar{X}_0 and \bar{X}_1 be the mean values of groups with $Y = 0$ and $Y = 1$, respectively.
 3. Let n_0 and n_1 be the number of cases in a group with $Y = 0$ and $Y = 1$, respectively, and S_X be the standard deviation of the random variable X .

Point Bi-Serial Correlation

The point bi-serial correlation is given by (Pearson, 1909 and Soper, 1914)

$$r_b = \frac{\bar{X}_1 - \bar{X}_0}{S_x} \sqrt{\frac{n_0 n_1}{n(n-1)}}$$

where n is the total number of cases in the sample and S_x is the standard deviation of X estimated from sample and is given by

$$S_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$$

Example

Ms. Sandra Ruth, data scientist at Airmobile, is interested in finding the correlation between the average call duration and gender of a person. The table provides the average call duration (measured in seconds) and gender of 30 customers of Airmobile. In the table , male is coded as 0 and Female is coded as 1. Calculate the point bi-serial correlation.

Solution

From the data we can calculate the following

$$\bar{X} = 345.33$$

$$\bar{X}_0 = 353.07$$

$$\bar{X}_1 = 339.4118$$

$$S_x = 71.7189$$

$$n_0 = 13$$

$$n_1 = 17$$

$$S_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$$

Point Bi-serial correlation is given by

$$r_b = \frac{\bar{X}_1 - \bar{X}_0}{S_x} \sqrt{\frac{n_0 n_1}{n(n-1)}} = \frac{339.4118 - 353.07}{71.7189} \sqrt{\frac{13 \times 17}{30(29)}} = -0.0960$$

There is a very low negative correlation between gender of a person and call duration.(A very low negative correlation of -0.0960 means that gender has a minimal influence, if any, on call duration. In practical terms, the gender of a person is not a good predictor of how long their phone calls will be.)

The Phi-Coefficient

Karl Pearson recommended the use of the **Phi-coefficient** when **both variables are binary** for calculating the association relationship (Cramer, 1946). Let X and Y be two random variables both taking binary values (that is, X takes values 0 or 1 and similarly Y also takes values either 0 or 1). One can create a contingency table as shown in table below.

	$Y = 0$	$Y = 1$	Total
$X = 0$	N_{00}	N_{01}	$N_{X0} = N_{00} + N_{01}$
$X = 1$	N_{10}	N_{11}	$N_{X1} = N_{10} + N_{11}$
Total	$N_{Y0} = N_{00} + N_{10}$	$N_{Y1} = N_{01} + N_{11}$	

Contingency table for presence or absence of two categorical variables

	Drink Coffee	Don't drink Coffee
Drink Tea	N_{11}	N_{10}
Don't drink Tea	N_{01}	N_{00}

The Phi-Coefficient

In the contingency table (Table in previous slide)

- N_{00} = Number of cases in the sample such that $X = 0$ and $Y = 0$
- N_{01} = Number of cases in the sample such that $X = 0$ and $Y = 1$
- N_{10} = Number of cases in the sample such that $X = 1$ and $Y = 0$
- N_{11} = Number of cases in the sample such that $X = 1$ and $Y = 1$
- N_{X0} = Number of cases in the sample such that $X = 0$
- N_{X1} = Number of cases in the sample such that $X = 1$
- N_{Y0} = Number of cases in the sample such that $Y = 0$
- N_{Y1} = Number of cases in the sample such that $Y = 1$
- The Phi-coefficient is given by

$$\varphi = \frac{N_{11}N_{00} - N_{10}N_{01}}{\sqrt{N_{X0}N_{X1}N_{Y0}N_{Y1}}}$$

Example

Joy Finance (JF) is a company that provides gold loans (in which gold is used as guarantee against the loan). Mr Georgekutty, Managing Director of JF, collected data to understand the relationship between loan default status (variable Y) and the marital status of the customer (variable X). Data is collected on past 40 loans and is shown in the table. Calculate the Phi-coefficient. In the table , $Y = 0$ implies non-defaulter, $Y = 1$ is defaulter, $X = 0$ is single, and $X = 1$ is married.

X	1	0	1	0	0	0	0	0	1	0
Y	0	1	0	1	0	0	0	1	1	1
X	0	1	1	0	0	1	0	0	0	1
Y	0	1	1	1	0	0	1	1	0	0
X	1	0	0	0	1	1	1	0	0	1
Y	0	0	0	1	0	0	0	0	0	0
X	1	0	0	0	1	0	1	0	1	1

Solution

The resultant contingency table is as follows

From the table

$$N_{00} = 13, N_{01} = 10, N_{10} = 10, N_{11} = 7$$

$$N_{X0} = 23, N_{X1} = 17, N_{Y0} = 23, N_{Y1} = 17$$

The Phi-coefficient is given by

		Y		Total
		0	1	
X	0	13	10	23
	1	10	7	17
Total		23	17	40

$$\phi = \frac{N_{11}N_{00} - N_{10}N_{01}}{\sqrt{N_{X0}N_{X1}N_{Y0}N_{Y1}}} = \frac{7 \times 13 - 10 \times 10}{\sqrt{23 \times 17 \times 23 \times 17}} = -0.0230$$

Since the Phi-coefficient is very small , we conclude that there is not much association between the marital status and loan default.

Test your understanding!

- In a study conducted, it was found that the number of suicides by hanging, strangulation or suffocation showed a positive increase with increase in US spending on science, space and technology. Does this mean that the two are related?
 - a) Yes, positive correlation means that the two variables are related
 - b) No, the variables are only related if there is negative correlation
 - c) No, this is an instance of spurious correlation
 - d) The information in this statement is not sufficient to arrive at a conclusion

Solution

c) No, this is an instance of spurious correlation

- Which correlation measure is applicable to data which is in ordinal, ratio or interval scale?

Solution

Spearman Rank Correlation

References

- Business Analytics by U. Dinesh Kumar – Wiley 2nd Edition, 2022
Chapters : 8.3 – 8.5



THANK YOU

Dr. Gowri Srinivasa

Professor, Department of Computer Science
and Engineering, PES University, Bengaluru

Email: gsrinivasa@pes.edu



DATA ANALYTICS

UE21CS342AA2

UNIT-1

**Lecture 1 : Introduction to DA , Data
Sources and Representations**

Gowri Srinivasa

Department of Computer Science and Engineering

Data Analytics

Unit 1

Lecture 12 : Hidden variables

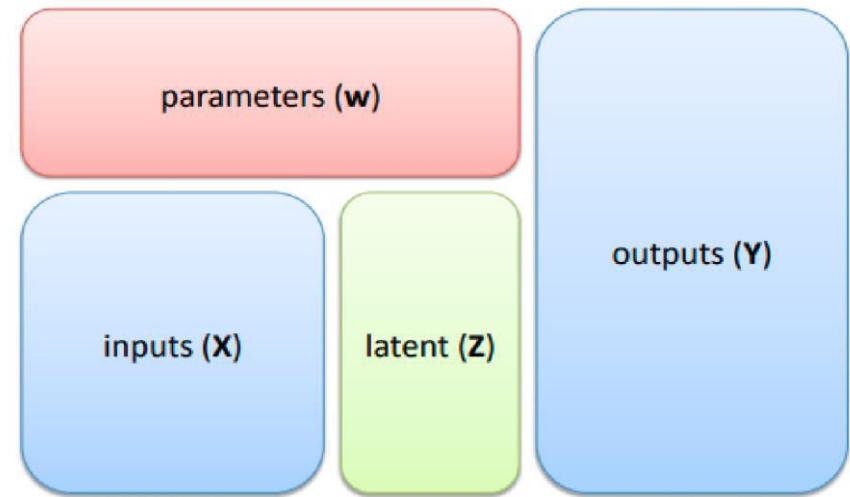
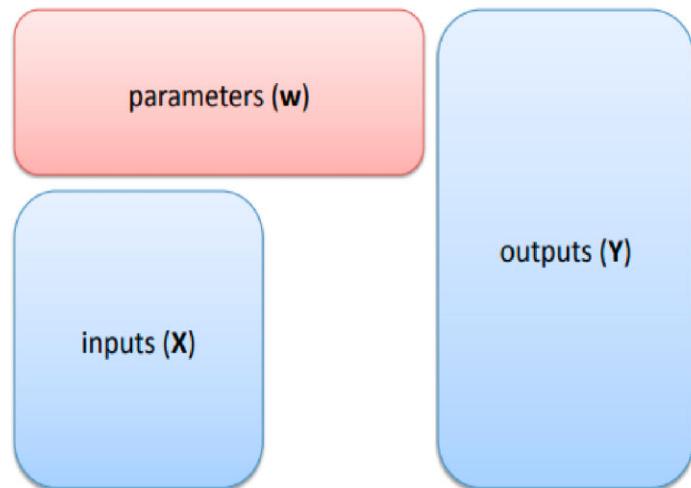
Gowri Srinivasa

Department of Computer Science and Engineering

With grateful thanks for contribution of slides to:
Prof. Swati Pratap Jagdale,
Professor at the Department of CSE, PESU

Hidden Variables

- Random variables in supervised learning Both input features and output labels can be represented as random variables. The goal of supervised learning is to model the relationship between the input features and the output labels.



- 'Hidden' (or latent) variable is one that we never 'see'
- Not even in training
- Sometimes we believe they are real
and sometimes they approximate reality (as it happens in Physics)

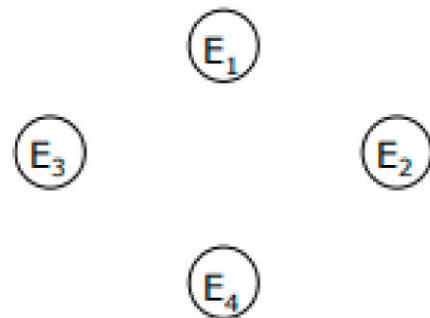
Learning With Hidden Variables

Why would we ever want to learn a hidden variables?

- One answer is: because we might be able to learn lower-complexity networks that way.
- Another is that sometimes such networks reveal interesting structure in our data.

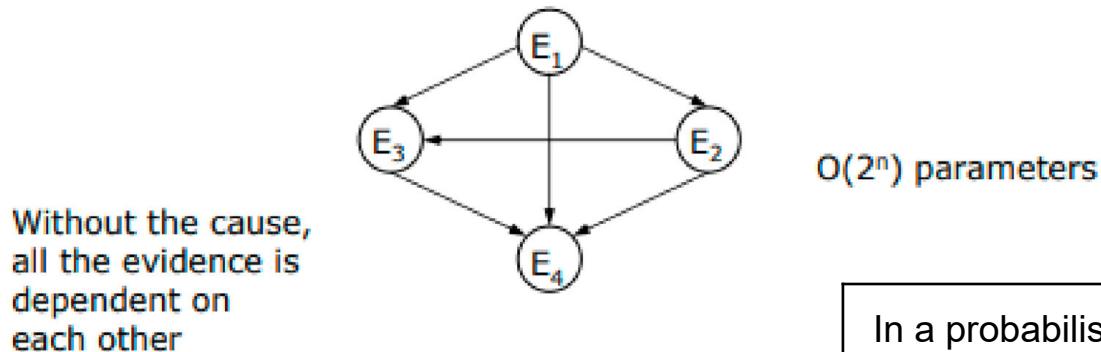
Learning With Hidden Variables

- Consider a situation in which you can observe a whole bunch of different evidence variables, E_1 through E_n . Maybe they are all the different symptoms that a patient might have. Or maybe they represent different movies and whether someone likes them.



Learning With Hidden Variables

- If those variables are all conditionally dependent on one another, then we would need a highly connected graph that is capable of representing the entire joint distribution between the variables.
- Because the last node has $n-1$ parents, it will take on the order of 2^n parameters to specify the conditional probability tables in this network.

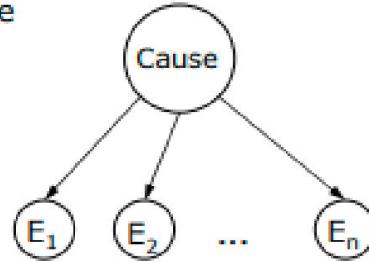


In a probabilistic graphical model, nodes represent variables, and edges represent probabilistic dependencies. If variables are conditionally dependent on one another, it means that the probability distribution of one variable is influenced by the values of other variables in the network.

Learning With Hidden Variables

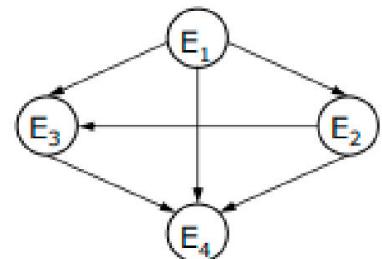
- But, in some cases, we can get a considerably simpler model by introducing an additional “cause” node.
- It might represent the underlying disease state that was causing the patients’ symptoms or some division of people into those who like drama and those who like comedies.

Cause is unobservable



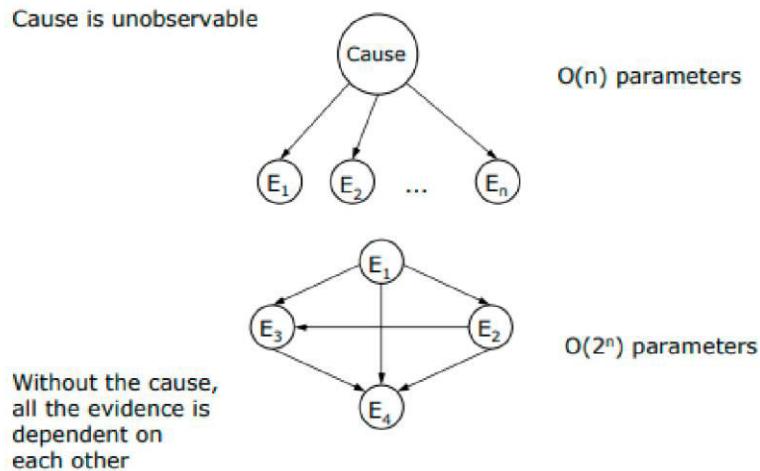
$O(2^n)$ parameters

Without the cause,
all the evidence is
dependent on
each other



Learning With Hidden Variables

- In the simpler model, the evidence variables are conditionally independent given the causes. That means that it would only require on the order of n parameters to describe all the CPTs in the network, because at each node, we just need a table of size 2 (if the cause is binary; or k if the cause can take on k values), and one (or $k-1$) parameter to specify the probability of the cause.



- So, what if you think there's a hidden cause? How can you learn a network with unobservable variables?

Simpson's Paradox

- Edward Hugh Simpson, a statistician and former cryptanalyst at Bletchley Park, described the statistical phenomenon - Simpson's paradox
- The art of data science is seeing beyond the data — using and developing methods and tools to get an idea of what that hidden reality looks like.
- Simpson's paradox showcases the importance of skepticism and interpreting data with respect to the real world, and also the dangers of oversimplifying a more complex truth by trying to see the whole story from a single data-viewpoint.

Simpson's Paradox

- *Simpson's Paradox:*

A trend or result that is present when data is put into groups that reverses or disappears when the data is combined.

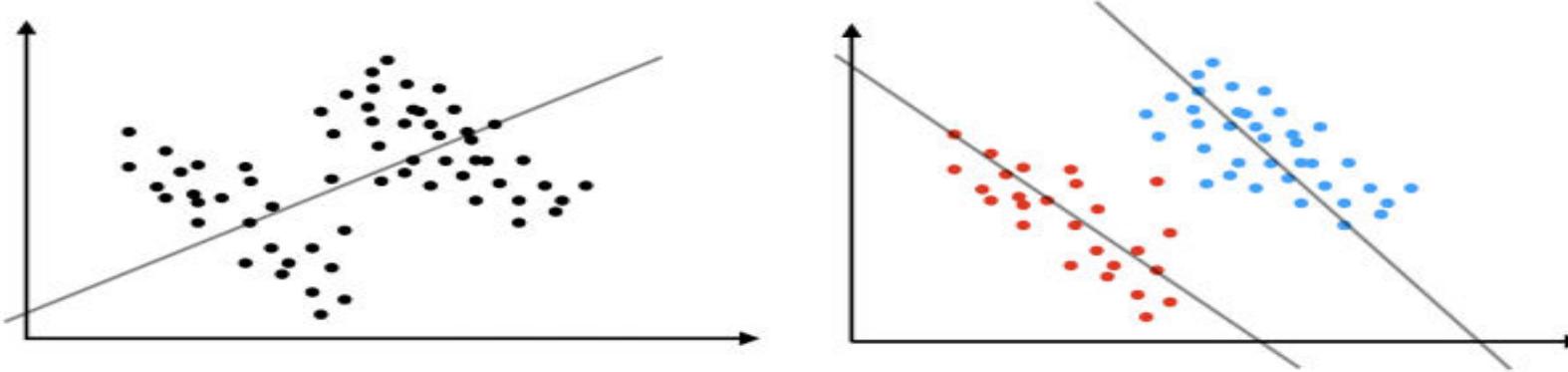
Example: UC Berkley's suspected gender-bias.

At the beginning of the academic year in 1973, UC Berkeley's graduate school had admitted roughly 44% of their male applicants and 35% of their female applicants.

Was there a discrimination against female applicants?

- When the data was studied department-wise, it was observed that:
- there was a statistically significant gender bias in favor of women for 4 out of the 6 departments, and no significant gender bias in the remaining 2
- It was discovered that women tended to apply to departments that admitted a smaller percentage of applicants overall, and that this hidden variable affected the marginal values for the percentage of accepted applicants in such a way as to reverse the trend that existed in the data as a whole

Simpson's Paradox



A visual example: the overall trend reverses when data is grouped by some colour-represented category.

Simpson's Paradox

- A simple example in business:
- Suppose the soft drinks industry is trying to choose between two new flavors they have produced. We could sample public opinion on the two flavors

Flavour	Sample Size	# Liked Flavour
Sinful Strawberry	1000	800
Passionate Peach	1000	750

80% of people enjoyed 'Sinful Strawberry' whereas only 75% of people enjoyed 'Passionate Peach'. So 'Sinful Strawberry' is more likely to be the preferred flavor.

Simpson's Paradox

- Some other information while conducting the survey, such as the gender of the person sampling the drink. What happens if we split our data up by gender?
- 84.4% of men and 40% of women liked 'Sinful Strawberry' whereas 85.7% of men and 50% of women liked 'Passionate Peach'

Flavour	# Men	# Liked Flavour (Men)	# Women	# Liked Flavour (Women)
Sinful Strawberry	900	760	100	40
Passionate Peach	700	600	300	150

Simpson's Paradox – An example

- According to our sample data, generally people prefer 'Sinful Strawberry', but both men and women separately prefer 'Passionate Peach'.
- This is an example of Simpson's Paradox!

Flavour	# Men	# Liked Flavour (Men)	# Women	# Liked Flavour (Women)
Sinful Strawberry	900	760	100	40
Passionate Peach	700	600	300	150

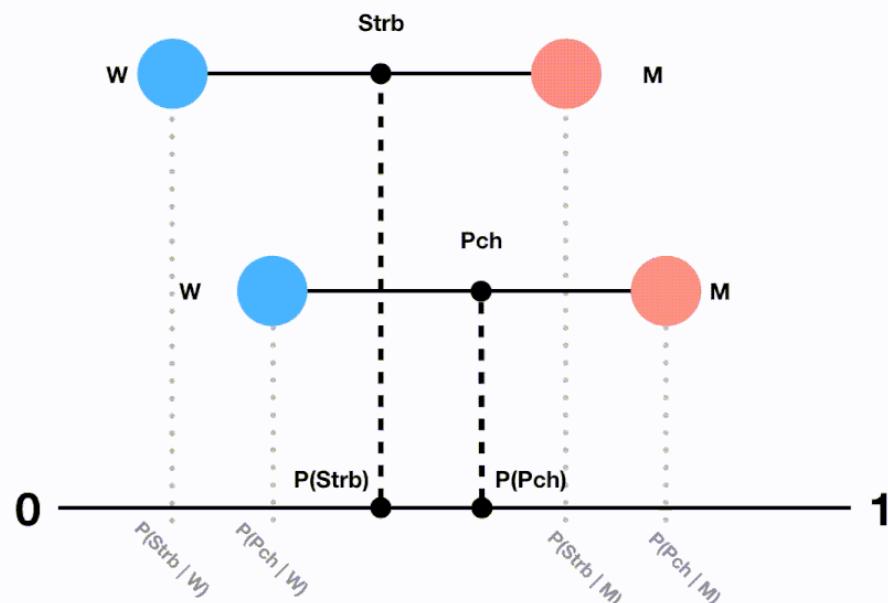
Simpson's Paradox

- Lurking variables (Hidden Variables)
- Simpson's paradox arises when there are hidden variables that split data into multiple separate distributions.
- Such a hidden variable is aptly referred to as a lurking variable, and they can often be difficult to identify.
- Consider the lurking variable (gender) and a little bit of probability theory:-
- $$\begin{aligned} P(\text{Liked Strawberry}) &= P(\text{Liked Strawberry} \mid \text{Man})P(\text{Man}) + \\ &P(\text{Liked Strawberry} \mid \text{Woman})P(\text{Woman}) \\ &= (760/900) \times (900/1000) + (40/100) \times (100/1000) = 800/1000 \end{aligned}$$
- $$\begin{aligned} P(\text{Liked Peach}) &= P(\text{Liked Peach} \mid \text{Man})P(\text{Man}) + P(\text{Liked} \\ &\text{Peach} \mid \text{Woman})P(\text{Woman}) \\ &= (600/700) \times (700/1000) + (150/300) \times (300/1000) = 750/1000 \end{aligned}$$

Flavour	# Men	# Liked Flavour (Men)	# Women	# Liked Flavour (Women)
Sinful Strawberry	900	760	100	40
Passionate Peach	700	600	300	150

Simpson's Paradox

- **Lurking variables (Hidden Variables)**
- We can think of the marginal probabilities of gender ($P(\text{Man})$ and $P(\text{Woman})$) as weights that, in the case of 'Sinful Strawberry', cause the total probability to be significantly shifted towards the male opinion.



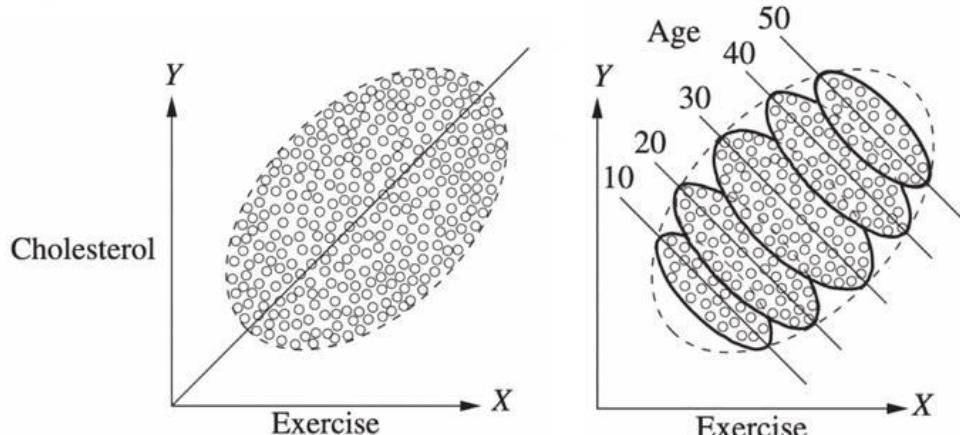
- Each coloured circle represents either the men or women that sampled each flavour, the position of the centre of each circle corresponds to that group's probability of liking the flavour.
- As the circles grow (i.e. sample proportions change) we can see how the marginal probability of liking the flavour changes.
- The marginal distributions shift and switch as samples become weighted with respect to the lurking variable (gender).

Causal inference

Causal Inference is the process where causes are inferred from data, from other variables in the model i.e., from observations or evidence.

How do hidden variables link to causal inference...?

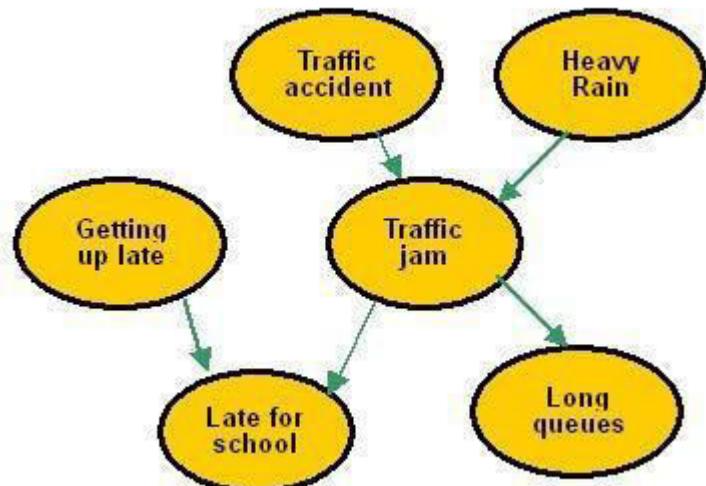
- Standard statistics is all about correlations, which are all good and fun, but correlations can lead to wrong assumptions



This is a graph showing the correlative relationship between Exercise and Cholesterol (which looks like a causal relationship but is not). If we just look at the correlative relationship between cholesterol and exercise, it looks like there's a causal relationship between the two. But this correlation actually happens because both cholesterol and exercise share a common cause or the confounder: age.

Confounder: hidden variable that relates to both the independent and dependent variable

Confounding variable – another example



- Independent variables:
 - Traffic accident, heavy rains, getting up late
- Dependent variables
 - Late for school, long queues
- Confounding variable
 - Traffic jam

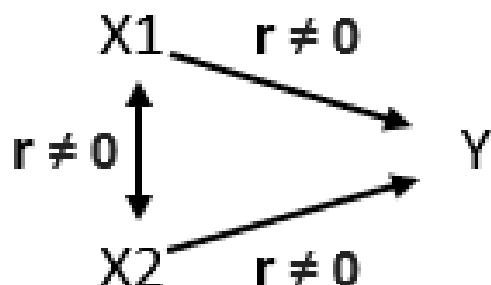
confounding variables

Confounding variables cause bias when they are omitted from the model. How can variables you leave out of the model affect the variables that you include in the model? At first glance, this problem might not make sense.

To be a confounding variable that can cause omitted variable bias, the following two conditions must exist:

- The confounding variable must correlate with the dependent variable.
- The confounding variable must correlate with at least one independent variable that is in the regression model.

Independent Dependent



This correlation structure causes confounding variables that are not in the model to bias the estimates that appear in your regression results

Why do we care about confounding variables?

Confounding Bias

- Bias is usually a result of errors in data collection or measurement.
- However, one definition of bias is “*...the tendency of a statistic to overestimate or underestimate a parameter*”, so in this sense, confounding is a type of bias.
- Confounding bias is the result of having confounding variables in your model. It has a direction, depending on if it over- or underestimates the effects of your model:
- **Positive confounding** is when the observed association is biased away from the null. In other words, it overestimates the effect.
- **Negative confounding** is when the observed association is biased toward the null. In other words, it underestimates the effect.

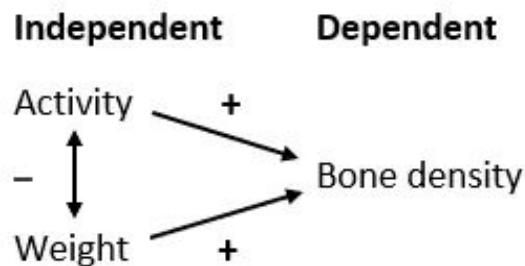
Confounding bias – another example

- Example of How Confounding Variables Can Produce Bias

Example:

- In a biomechanics lab, one study assessed the effects of physical activity on bone density.
- They measured various characteristics including the subjects' activity levels, their weights, and bone densities among many others.
- Theories about how our bodies build bone suggest that there should be a positive correlation between activity level and bone density. In other words, higher activity produces greater bone density.
- Simple regression analysis to determine whether there is a relationship between activity and bone density... **there was no relationship at all!**

- They included activity level as the only independent variable, but it turns out there is another variable that correlates with both activity and bone density—the **subject's weight**.

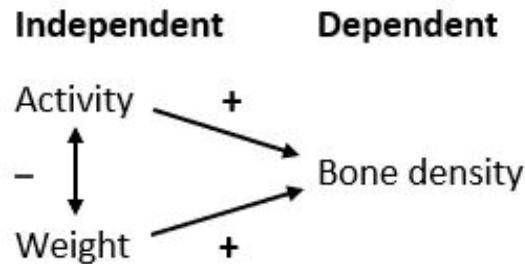


The diagram shows the signs of the correlations between the variables.

- Subjects who are more active tend to have higher bone density. Additionally, subjects who weigh more also tend to have higher bone density. However, there is a negative correlation between activity and weight. More active subjects tend to weigh less.

This correlation structure produces two opposing effects of activity. More active subjects get a bone density boost. However, they also tend to weigh less, which reduces bone density.

Confounding bias – another example



A model that does not account for ‘weight’ but only uses ‘activity’ ends up with a negative bias: the model underestimates the importance of activity

In other words, without ‘weight’, ‘activity’ is deemed ‘not important’ for ‘bone density’ (as it has opposing effects)!

The diagram shows the signs of the correlations between the variables.

In this study ‘weight’ is the confounding variable.

If we try to build a model that predicts bone density based only on ‘activity’, then,

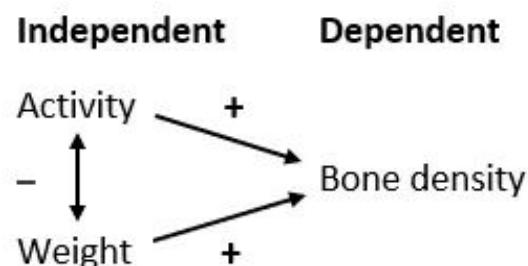
More activity → higher bone density
(for those who are more active)

More activity → lower bone density
(for those who weigh less)

Two opposing effects have to be attributed to one variable: ‘activity’

Confounding bias – another example

	Included and Omitted: Negative Correlation	Included and Omitted: Positive Correlation
Included and Dependent: Negative Correlation	Positive bias: coefficient is overestimated.	Negative bias: coefficient is underestimated.
Included and Dependent: Positive Correlation	Negative bias: coefficient is underestimated.	Positive bias: coefficient is overestimated.



Included (activity) and omitted (weight) are negatively correlated.
The included variable (activity) and the dependent variable (bone density)
have a positive relationship, implies the result has a **negative bias**.

How do we identify a confounding variable?

- Confounding variables cause omitted variable bias
 - They are also called confounders
- A confounding variable is closely related to both the independent and dependent variables in a study. An independent variable represents the suppose *cause*, while the dependent variable is the supposed *effect*.
- A confounding variable is a third variable that influences both the independent and dependent variables. Failing to account for confounding variables can cause you to wrongly estimate the relationship between your independent and dependent variables.
- **A lurking(HIDDEN) variable is one that was simply not included in the study**
- How do we identify a confounding variable?
 1. There must be three or more variables in the study
(two variables => one is the cause the other is the effect)
 2. The variable we suspect is a confounding variable, changes systematically with at least one of the variables we are measuring (either independent or dependent)
 3. Identify extraneous variables that relate to subjects (age, gender, etc.), the environment in which the study is conducted (weather, location, etc.) and to the two variables we are explicitly measuring to test for systematic changes to identify a confounding variable.

How do we reduce the effects of a confounding variable?

- Make sure you identify all of the possible confounding variables in your study.
- Make a list of everything you can think of and one by one, consider whether those listed items might influence the outcome of your study. Usually, someone has done a similar study before you. So check the academic databases for ideas about what to include on your list.
- Once you have figured out the variables, techniques to reduce the effect of those confounding variables:
 - Bias can be eliminated with random samples.
 - Introduce control variables to control for confounding variables. For example, you could control for age by only measuring 30 year olds.
 - Within subjects designs test the same subjects each time. Anything could happen to the test subject in the “between” period so this does not make for perfect immunity from confounding variables.
 - Counterbalancing can be used if you have paired designs.

In counterbalancing, half of the group is measured under condition 1 and half is measured under condition 2.

How do we reduce the effects of a confounding variable?

- Random sampling is a technique where you select participants from a population in a way that each individual has an equal chance of being included in the sample. This helps to minimize selection bias and ensures that the sample is representative of the population.
- Control variables are variables that you include in your analysis to account for their potential influence on the relationship you are investigating. By statistically controlling for these variables, you can isolate the relationship between the variables of interest and reduce the impact of confounding. For example, if you are studying the relationship between a treatment and an outcome, controlling for age could help minimize the influence of age-related confounding.
- Within-Subjects Designs: A within-subjects design involves testing the same group of participants under different conditions. By doing this, each participant serves as their control, reducing the impact of individual differences that might otherwise confound the results. However, as you mentioned, this approach doesn't completely eliminate all potential confounding variables, particularly those that can change between measurement occasions.
- Counterbalancing:
 - Counterbalancing is a technique often used in experimental designs, especially when studying paired conditions. It involves ensuring that each condition is presented to different participants in a systematic order. This helps to minimize the potential impact of order effects (e.g., learning or fatigue) as confounding variables by distributing their influence across different groups of participants.

Hidden Variable:

- A hidden variable, also known as a latent variable, is a variable that is not directly measured or observed in a study but is believed to have an impact on the variables of interest.
- Not a part of the study
- Hidden variables are often used in statistical models to account for unobservable factors that might be influencing the observed data.

confounding Variable:

- A confounding variable is a variable that is both related to the independent variable (the variable you are studying) and the dependent variable (the outcome you are interested in), making it appear as if there is a cause-and-effect relationship between the independent and dependent variables when, in fact, there isn't.
- Part of the study
- Confounding variables can lead to incorrect or misleading conclusions in research.

References

- <http://www.cs.cmu.edu/~nasmith/psnlp/lecture5.pdf>
- <https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-825-techniques-in-artificial-intelligence-sma-5504-fall-2002/lecture-notes/Lecture18FinalPart1.pdf>
- <https://towardsdatascience.com/simpsons-paradox-and-interpreting-data-6a0443516765>
- <https://www.statisticshowto.com/experimental-design/confounding-variable/>
- <https://statisticsbyjim.com/regression/confounding-variables-bias/>

Test your understanding

- It is observed that ice cream sales increase in the summer and number of robberies also increase in the summer. It is later discovered that 'summer' (high temperature) is when people have holidays and travel, hence their homes are vacant. Summer is also when ice cream sales increase. Is 'summer' (temperature) a hidden (lurking) variable in this study or a confounding one?

Hidden or lurking

(as it was not a part of the study)

- While studying the effect of sunlight on Vitamin D levels versus dietary supplements on the level of Vitamin D, group A subject to sunlight had a lot of orange juice (infused with Vitamin D). Is 'orange juice (infused with Vitamin D)' a hidden variable or a confounding variable in this study?

Confounding

(it was a part of the study, but does not allow us to isolate the effect of sunlight alone versus dietary supplement – is the increase in the level of vitamin D due to orange juice or sunlight? We need further studies to infer the effect of sunlight alone).

References

<http://www.cs.cmu.edu/~nasmith/psnlp/lecture5.pdf>

<https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-825-techniques-in-artificial-intelligence-sma-5504-fall-2002/lecture-notes/Lecture18FinalPart1.pdf>

<https://towardsdatascience.com/simpsons-paradox-and-interpreting-data-6a0443516765>

<https://www.statisticshowto.com/experimental-design/confounding-variable/>

<https://statisticsbyjim.com/regression/confounding-variables-bias/>



PES
UNIVERSITY

THANK YOU

Dr. Gowri Srinivasa

Professor, Department of Computer Science
and Engineering, PES University, Bengaluru

Email: gsrinivasa@pes.edu



DATA ANALYTICS

UE21CS342AA2

UNIT-1

**Lecture 2 : The R programming
environment and descriptive statistics**

Gowri Srinivasa

Department of Computer Science and Engineering

Data Analytics

Unit 1

Lecture 2 : The R programming environment and descriptive statistics.

Slides excerpted from: U. Dinesh Kumar,
“Business Analytics”, Wiley, 2nd Edition 2022

Gowri Srinivasa
Department of Computer Science and Engineering

Slides collated by:

Nishanth M S PESU-2023, Department of CSE
nishanthmsathish.23@gmail.com

Harshitha Srikanth ,VII Sem ,PESU,Department of CSE
harshithasrikanth13@gmail.com

With grateful thanks for contribution of slides to:
Dr. Mamatha H R, Professor at the Department of CSE, PESU

What is R?

- **Free, open source interpreted language**
Ross Ihaka & Robert Gentleman in 1993.
- **Most widely used data analysis software**
Used by 2M+ data scientists, statisticians and analysts
- **Most powerful statistical programming language**
Flexible, extensible and comprehensive for productivity
- **Create beautiful and unique data visualizations**
As seen in New York Times, Twitter and Flowing Data
- **Thriving open-source community**
Leading edge of analytics research



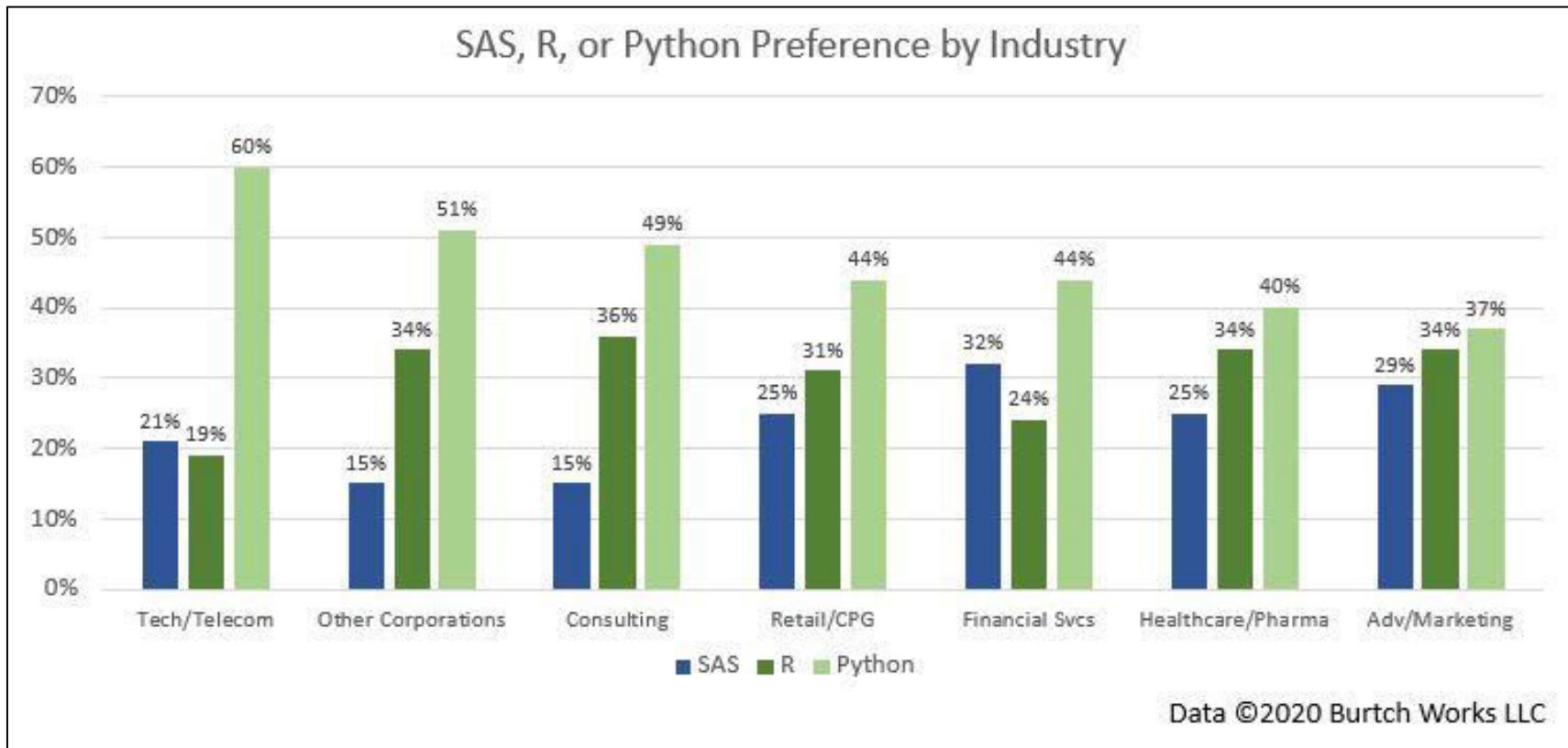
- R allows to collect data in real-time, perform statistical and predictive analysis, create visualizations and communicate actionable results to stakeholders
- It houses more than 9100 packages of statistical functions.
- R's expressive syntax allows to quickly import, clean and analyze data from various data sources.
- R also has charting capabilities, which means you can plot your data and create interesting visualizations from any dataset.

- R is used in predictive analytics and machine learning.
- Packages for ML tasks like linear and non-linear regression, decision trees, linear and non-linear classification and many more.
- R has been used primarily in academics and research and is great for exploratory data analysis.
- In recent years, enterprise usage has rapidly expanded. It is used by statisticians, engineers, and scientists without computer programming skills.
- It is popular in academia, finance, pharmaceuticals, media, and marketing.

Data Handling Capabilities

- R is great for data analysis because of its huge number of packages, readily usable tests, and the advantage of using formulas.
- It can handle basic data analysis without needing to install packages.
- Big datasets require the use of packages such as data.table and dplyr.

Preference by Industry



- The **S** language has been developed since the late 1970s by **John Chambers** and colleagues at Bell Labs as a **language for programming with data**.
- S language combines ideas from a variety sources (awk, lisp, APL,) and provides an environment for quantitative computations and visualization.
- Provides an explicit and consistent structure for manipulating, analyzing statistically and visualizing data.
- **S-Plus** is a commercialization of the Bell Labs framework. It is “S” plus “graphics”.
- R is a free implementation of a dialect of [the S language](#)

- R is an Open source statistical environment/platform developed by Robert Gentleman and Ross Ihaka (University of Auckland) during the 1990s.
- Currently maintained by the R core-development team, a hard-working, international team of volunteer developers.
- [The primary R system is available from the Comprehensive R Archive Network, also known as CRAN.](#)
- CRAN also hosts many add-on packages that can be used to extend the functionality of R. Over 6,789 packages are available on CRAN that have been developed by users and programmers around the world.

Finding out the latest version of R:

- To find out what is the latest version of R, you can look at [the CRAN \(Comprehensive R Network\) website, http://cran.r-project.org/.](http://cran.r-project.org/)

Installing R on Windows:

- To install R on your Windows computer, follow the below steps :
- Go to <https://cran.rstudio.com/bin/windows/base/>
- Alternative : <http://ftp.heanet.ie/mirrors/cran.r-project.org>
- Under “Download and Install R”, click on the “Windows” link.
- Download the required.exe file. The latest version of R is 4.3.1. Click on the link which says “Download R 4.3.1 for windows.”
- After downloading, double click on the executable to run it.
- Choose English as the language to install it.

Download and install RStudio on Windows

Visit <https://www.rstudio.com/products/rstudio/download/>
and click on DOWNLOAD RSTUDIO DESKTOP

List of IDEs

- RStudio
- StatET for R (eclipse based)
- R-Brain IDE (RIDE)
- IntelliJ IDEA
- R Tool for Visual Studio

R getwd() Function

- Working directory is the directory where R finds all R files for reading and writing.
 - `getwd()` function returns an absolute file path representing the current working directory of the R process.
 - `getwd()` **Output:"C:/Users/******/Documents"**

R setwd() Function

- `setwd(dir)` is used to set the working directory to `dir`.
 - `setwd("e:/folder")` Output: Sets `pwd` to "e:/folder"

Setting your Working Directory

R dir() Function

- `dir()` function lists all the files in a directory.
- **dir()** **Output:Lists all files in the pwd**

R ls() Function

- `ls()` is a function in R that lists all the objects in the working environment.
- **ls()** **Output:Lists all objects in the pwd**
- It can be used in the scenario where you want to clean the environment before running the code. The following command will remove all the object from R environment.
- **rm(list = ls())**

Getting help with R

- To get general help just type the below command
- **help.start()**
- To access documentation for the standard lm (linear model) function, for example, enter the command.
- **help(lm) / help("lm") / ?lm /?"lm"**
- To see the list of pre-loaded data (datasets), type the function:
- **data()**
- A guide to get started on working with built in datasets:

Installing an R Package in RStudio

- The primary location for obtaining R packages is [CRAN](#).
- Information about the available packages on CRAN with the ***available.packages()*** function.
- **a <- available.packages()**
- Packages can be installed with the ***install.packages()*** function in R.
- **install.packages("ggplot2")**
- Install multiple R packages at once with a single call to *install.packages()*. Place the names of the R packages in a character vector.
- **install.packages(c("caret","ggplot2","dplyr"))**

Installing an R Package in RStudio



Loading R Packages

- Installing a package does not make it immediately available to you in R; you must load the package. The `library()` function is used to load packages into R.
- **library(ggplot2)**
- After loading a package, the functions exported by that package will be attached to the top of the `search()` list (after the workspace).
- **library(ggplot2)**
- **search()**
- To save your workspace to a file, you may type `save.image()` or use **Save Workspace** in the **File** menu
- The default workspace file is called **.RData**



THANK YOU

Dr. Gowri Srinivasa

Professor, Department of Computer Science
and Engineering, PES University, Bengaluru

Email: gsrinivasa@pes.edu

PES University-EC Campus
Data Analytics
Faculty : Dr R Bharathi
Tutorial-1

Exercise1: Use the two methods below to normalize the following group of data: 200, 300, 400, 600, 1000 .

- (a) min-max normalization by setting $\min = 0$ and $\max = 1$ (b) z-score normalization

Answer 1:

Answer:

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

<i>original data</i>	200	300	400	600	1000
<i>[0,1] normalized</i>	0	0.125	0.25	0.5	1

- (b) z-score normalization

<i>original data</i>	200	300	400	600	1000
<i>z-score</i>	-1.06	-0.7	-0.35	0.35	1.78

Exercise-2: Suppose a group of 12 sales price records has been sorted as follows:

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

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

- (a) equal-frequency partitioning (b) equal-width partitioning

Answer:

- (a) equal-frequency partitioning

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

- (b) equal-width partitioning

The width of each interval is $(215 - 5)/3 = 70$.

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

Exercise:3
Page number:222(Business Analytics- I edition)

Harrison Seth, Dean of a Business School, believes that the outgoing salary of their MBA students may be correlated with their undergraduate specialization. Harrison believes

PES University-EC Campus
Data Analytics
Faculty : Dr R Bharathi
Tutorial-1

that the students with engineering specialization at the undergraduate degree received more salary compared to other degrees. Table 8.14 shows the outgoing salary (in millions of rupees) of MBA graduates and their discipline in undergraduate (1 = engineering and 0 = non-engineering). Calculate the correlation between salary and engineering discipline,

TABLE 8.14 Salary (in millions of rupees) and undergraduate degree

	(1 = engineering and 0 = non-engineering)									
Degree	0	1	0	1	0	0	1	0	0	1
Salary	3.3	2.22	1.82	2.55	1.84	2.53	2.87	2.39	2.32	2.79
Degree	1	1	0	1	0	0	1	1	0	0
Salary	2.22	2.31	2.05	2.04	1.7	2.28	2.56	3.13	2.26	2.56
Degree	0	0	0	0	1	0	0	0	1	1
Salary	2.03	1.45	1.62	0.92	2.31	2.37	1.59	2.56	3.13	3

[Answer 3: Refer Solution manual](#)

Exercise4.Two-Way ANOVA

The hypothesis tests associated with two-way ANOVA are as follows:

1. Test of Factor A Main Effects:

$$H_0: \alpha_i = 0 \text{ for all } i (i = 1, 2, \dots, a)$$

$$H_A: \text{Not all } \alpha_i \text{ are zero}$$

2. Test of Factor B Main Effects:

$$H_0: \beta_j = 0 \text{ for all } j (j = 1, 2, \dots, b)$$

$$H_A: \text{Not all } \beta_j \text{ are zero}$$

3. Test of Interaction Effects:

$$H_0: \alpha_i\beta_j = 0 \text{ for all } i (i = 1, 2, \dots, a) \text{ and } j (j = 1, 2, \dots, b)$$

$$H_A: \text{Not all } \alpha_i\beta_j \text{ are zero}$$

Numerical example

A two-way ANOVA is used to determine whether or not there is a statistically significant difference between the means of three or more independent groups that have been split on two factors.

Suppose a gardener wants to know if **plant growth** is influenced by **sunlight exposure** and **watering frequency**. She plants 40 seeds and lets them grow for one month under different conditions for **sunlight exposure** and **watering frequency**(**Independent Variables**).

PES University-EC Campus
Data Analytics
Faculty : Dr R Bharathi
Tutorial-1

In the table.1 below, there are five plants grown under each combination of conditions. For example, there were five plants grown with daily watering and no sunlight and their heights after two months are 4.8 inches, 4.4 inches, 3.2 inches, 3.9 inches, and 4.4 inches. After one month, **the heights(Dependent Variable)** are recorded for each plant as shown in Table.1.

Table 1

Watering Frequency	Sun light Exposure			
	No light	Low	Medium	High
DAILY	4.8	5	6.4	6.3
	4.4	5.2	6.2	6.4
	3.2	5.6	4.7	5.6
	3.9	4.3	5.5	4.8
	4.4	4.8	5.8	5.8
WEEKLY	4.4	4.9	5.8	6
	4.2	5.3	6.2	4.9
	3.8	5.7	6.3	4.6
	3.7	5.4	6.5	5.6
	3.9	4.8	5.5	5.5

The sum of squares in the case of two-way ANOVA with equal sample sizes is given by

$$SST = SSA + SSB + SSAB + SSW \quad \text{-----} (Eq-a)$$

- A- First Factor (Watering Frequency)
B- Second Factor (Sunlight Exposure)

- a- 2(Number of groups in Factor A- Watering frequency)
- b- 4(Number of groups in Factor B- Sunlight Exposure)
- c- 5(Number of observations in each group-ASSUMED TO BE SAME FOR ALL GROUPS)

Step 1: Calculation of All “mean”

Watering Frequency	Sun light Exposure			
	No light	Low	Medium	High
DAILY	4.8	5	6.4	6.3
	4.4	5.2	6.2	6.4
	3.2	5.6	4.7	5.6
	3.9	4.3	5.5	4.8
	4.4	4.8	5.8	5.8
MEAN	4.14			
Mean of Daily = $(4.8 + 5 + 6.4 + 6.3 + \dots + 4.4 + 4.8 + 5.8 + 5.8) / 20$				
= 5.155				
WEEKLY	4.4	4.9	5.8	6
	4.2	5.3	6.2	4.9

PES University-EC Campus
Data Analytics
Faculty : Dr R Bharathi
Tutorial-1

	3.8	5.7	6.3	4.6
	3.7	5.4	6.5	5.6
	3.9	4.8	5.5	5.5
Mean of Weekly = $(4.4 + 4.9 + 5.8 + 6 + \dots + 3.9 + 4.8 + 5.5 + 5.5) / 20$ = 5.15				
MEAN(BOTH)	4.07	5.1	5.89	5.55
Grand mean = $(4.8 + 5 + 6.4 + 6.3 + \dots + 3.9 + 4.8 + 5.5 + 5.5) / 40$ = 5.1525				
the mean height of all plants watered daily: Mean of Daily = $(4.8 + 5 + 6.4 + 6.3 + \dots + 4.4 + 4.8 + 5.8 + 5.8) / 20$ = 5.155				
the mean height of all plants watered weekly: Mean of Weekly = $(4.4 + 4.9 + 5.8 + 6 + \dots + 3.9 + 4.8 + 5.5 + 5.5) / 20$ = 5.15				

Step 2: Calculate Sum of Squares for First Factor A (Watering Frequency)

$$SSA = b \times c \times \sum_{i=1}^a (\mu_i - \mu)^2$$

where μ_i is the mean of all observations in level i of factor A and c is the number of observations in each group (assumed to be same for all groups).

$$SSA = 20(5.155 - 5.1525)^2 + 20(5.15 - 5.1525)^2 = .00025 \text{ ----- (Eq1)}$$

Step 3: Calculate Sum of Squares for Second Factor B (Sunlight Exposure)

$$SSB = a \times c \times \sum_{j=1}^b (\mu_j - \mu)^2$$

Here μ_j is the mean of all observations in level j of factor B.

$$SSB = 10(4.07 - 5.1525)^2 + 10(5.1 - 5.1525)^2 + 10(5.89 - 5.1525)^2 + 10(5.55 - 5.1525)^2 = 18.76475 \text{ ----- (Eq2)}$$

Step 4: Calculate Squares of Deviation(error)Within groups(SSW)

$$SSW = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^c (Y_{ijk} - \mu_{ij})^2$$

Next, we will calculate the sum of squares within by taking the sum of squared differences between each combination of factors and the individual plant heights.

For example, the mean height of all plants watered daily with no sunlight exposure is 4.14. We can then calculate the sum of squared differences for each of these individual plants as:

PES University-EC Campus
Data Analytics
Faculty : Dr R Bharathi
Tutorial-1

- SS for daily watering and no sunlight: $(4.8-4.14)^2 + (4.4-4.14)^2 + (3.2-4.14)^2 + (3.9-4.14)^2 + (4.4-4.14)^2 = \mathbf{1.512}$
- SS for daily watering and low sunlight: **0.928**
- SS for daily watering and medium sunlight: **1.788**
- SS for daily watering and high sunlight: **1.648**
- SS for weekly watering and no sunlight: **0.34**
- SS for weekly watering and low sunlight: **0.548**
- SS for weekly watering and medium sunlight: **0.652**
- SS for weekly watering and high sunlight: **1.268**

Sums of squares within = $1.512 + .928 + 1.788 + 1.648 + .34 + .548 + .652 + 1.268 = \mathbf{8.684}$ --
-----**(Eq3)**

Step 5: Calculate Total Sum of Squares

$$SST = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^c (Y_{ijk} - \mu)^2$$

where c is the number of observations in each group and μ is the overall mean.

Next, we can calculate the total sum of squares by taking the sum of the differences between each individual plant height and the grand mean:

Total Sum of Squares = $(4.8 - 5.1525)^2 + (5 - 5.1525)^2 + \dots + (5.5 - 5.1525)^2 = \mathbf{28.45975}$ --
-----**(Eq4)**

Step 6: Calculate Sum of Squares Interaction

$$SSAB = c \times \sum_{i=1}^a \sum_{j=1}^b (\mu_{ij} - \mu_i - \mu_j - \mu)^2$$

where μ_{ij} is the average of i^{th} level of factor A and j^{th} level of factor B.

But we know that from (Eq-a),

$$\begin{aligned} SST &= SSA + SSB + SSAB + SSW \\ SSAB &= SST - SSA - SSB - SSW \end{aligned}$$

$$\begin{aligned} SSAB &= 28.45975 - .00025 - 18.76475 - 8.684 \\ SSAB &= \mathbf{1.01075} \end{aligned}$$

Step 7: Fill in 2 WAY-ANOVA Table

PES University-EC Campus
Data Analytics
Faculty : Dr R Bharathi
Tutorial-1

Sum of squares of deviation for various effects and the corresponding F-statistic in a two-way ANOVA with equal sample size:

Sum of Squared Variation	Degrees of Freedom	Mean Squared Variation	F-Statistics
SSA	a – 1	$MSA = SSA/(a - 1)$	$F = MSA/MSW$
SSB	b – 1	$MSB = SSB/(b - 1)$	$F = MSB/MSW$
SSAB	(a – 1)(b – 1)	$MSAB = SSAB/(a - 1)(b - 1)$	$F = MSAB/MSW$
SSW	ab(c – 1)	$MSW = SSW/ab(c - 1)$	

Let us fill the values in our 2-way ANOVA table: alpha= 0.05

Sum of Squared Variation	Degrees of Freedom	Mean Squared Variation	F-Statistics	F-Critical	p-value
SSA(watering frequency) =0.00025	2 – 1=1	$MSA = 0.00025/1=0.00025$	$F = MSA/MSW=0.000921$	$=(DF_{ssa},DF_{within})=4.14910$	0.975
SSB(Sunlight exposure)=18.76475	4 – 1=3	$MSB = 18.76475/3=6.254917$	$F = MSB/MSW=23.04898$	$F(3,32)=0$	<.000
SSAB (Interaction)=1.01075	(2 – 1)(4 – 1)=3	$MSAB = 1.01075/3=0.336917$	$F = MSAB/MSW=1.241517$	$F(3,32) = 0$	0.311
SSW(within)=8.684	2*4(5 – 1)=32	$MSW = 8.684/32=0.271375$			
SST		28.45975			

Step 8: Interpret the results

We can observe the following from the ANOVA table:

- The p-value for the interaction between watering frequency and sunlight exposure was **0.311**. This is not statistically significant at $\alpha = 0.05$. Hence
- The p-value for watering frequency was **0.975**. This is not statistically significant at $\alpha = 0.05$.
- The p-value for sunlight exposure was < **0.000**. This is statistically significant at $\alpha = 0.05$.

If p-value < 0.05 – “significant”- reject the null hypothesis
 p-value>0.05 – not statistically significant- retain the null hypothesis