classmate
Date
Page

* NOTE *

ii) I guess, these are complete notes, but still go through Textbook once again

iii) Practice Numericals in Unit II, Solve Numerical for chi-Square Test, I forgot about it :c

PEACE 

UNIT- I

Introduction To Data Science & Big Data.

- **Definition -**

  i) **Data :**

  - It is collection of facts & figures which relay something specific, but is not organized in any way.
  - It can be numbers, words, observations, etc.
  - Data can be said as raw material in production of information.

* Data Science & Big Data -

→

- Data science is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structural and unstructured data.

- It requires a variety of tools to extract information from data.

- Principal purpose of Data Science is to find patterns within Data.

- It uses various statistical techniques to analyze and draw insights from data

- Applications -
  i) Fraud & Risk Detection
  ii) Healthcare
  iii) Internet Search
  iv) Website Recommendations
  v) Speech Recognition
  vi) Image Recognition
  vii) Airline Route Planning

- Big Data -
  - It is a field that treats ways to analyze, systematically extract information from, or deal with datasets that are too large or complex to be dealt with using same software.
  - Big data can be analyzed for insights that lead to better decisions & strategic buisness moves.
  - Advantages -
    i) Product Development
    ii) Manufacturing
    iii) Marketing
    iv) Price Management
  - Process of big data begins with raw data that isn't aggregated / organized & is often impossible to store in memory of single computer.

★ 5 V's of Big Data -

→ 1] Volume :
- It refers to sheer size or quantity of data generated or collected
- Eg, Terabytes, petabytes of data

2] Velocity :
- The term 'Velocity' relates to speed at which data is generated, processed & updated
- Eg, Real-time data processing

3] Variety :
- It refers to diverse types & formats of data, including structured and unstructured
- Eg, text, images, videos, etc.

4] Value :
- It represents the importance of deriving meaningful insights & value from data
- Eg, Extracting actionable information

5] Veracity :
- It refers to accuracy, reliability & trustworthiness of data
- Eg, Dealing with uncertainities, noise in data

★ Comparison -

Data Science	Business Intelligence
i) It focuses on extracting insights and making predictions	i) It focuses on analyzing past & present data for business insight
ii) Purpose of DS is to predict future trends and behaviours	ii) Purpose of BI is to support decision-making process within organizations
iii) Techniques used in DS are ~~Data collection, organization & visualization~~ Statistical Analysis, ML, Modeling	iii) Techniques used in BI are Data collection, organization & visualization
iv) DS can handle both structured & unstructured data	iv) BI is designed to handle static & highly structured data

★ Data Types in DS -

→ • Data can be divided in 2 types:

1) Qualitative Data -

• It provides information about quality of an object or information which cannot be measured

ⓐ Nominal Data:
- Nominal data usually deals with non-numeric variables or numbers
- Eg, Gender: male, Female, Other

ⓑ Ordinal Data:
- Ordinal data is variable in which value of data is captured from ordered set
- Eg, University ranking: 1st, 9th, 87th, ...

2) Quantitative Data -

• It is the one that focuses on numbers & mathematical calculations & can be computed.

ⓐ Interval Data:
- It corresponds to a variable in which value is chosen from interval set
- Eg, Celsius Temperature

⑥ Ratio Data:

- Any variable for which ratios can be computed & are meaningful is called ratio data.
- Eg, Age, Weight, Height, etc.

★ Data Wrangling -

→ • Data wrangling is the process of getting data from its raw format into something suitable for more conventional analytics.

• The goal of data wrangling is to assure quality and useful data.

• Steps in Data wrangling are:

1) Discovering -
   • The first step of data wrangling is to gain better understanding of data

2) Structuring -
   • The next step is to organize the data

3) Cleaning -
   • Null values in data are changed and standard formatting is implemented

4) Enriching -

- At this step, we determine whether or not additional data would benefit the data set that could be easily added

5) Validating -

- Validation rules are repetitive programming sequences that verify data consistency, quality & security

6) Publishing -

- Prepare the data set for use downstream, which could include use of users or software.

★ Data Integration -

→ • Data integration combines data from multiple sources to form a coherent data store.

• Metadata, data conflict detection & correlation analysis contribute toward smooth data integration

• Data integration is important as it provides unified view of scattered data & maintains accuracy.

• Issues in Data Integration:
i) Entity Identification Problem
ii) Redundancy

★ Data Transformation -

→ • In data transformation, data are transformed or consolidated into forms appropriate for mining.

• Data transformation involves following:
1) Smoothing:
It removes noise from data using binning, regression

2) Aggregation:

An aggregation or summary operation is applied to data

3) Generalization:

Low-level data is replaced by higher-level concepts

4) Normalization:

Attribute data are scaled, so as to fall within specified range

5) Attribute construction:

New attributes are constructed and added from given set of attributes.

★ Data Reduction -

→ • Data Reduction is nothing but obtaining a reduced representation of data set that is smaller in volume yet produces same analytical results.

• It is the process that reduces volume of original data & represents it in much smaller volume.

• Data reduction techniques ensure integrity of data while reducing data

▪ methods :

1) Dimensionality Reduction -

• It is the process of reducing number of random variables under consideration, by obtaining set of principle variables.

2) Clustering -

• Cluster algorithm can be applied to discretize a numerical attribute (A) by partitioning the values of A into clusters or groups

3) Sampling -

• Sampling can be used as data reduction technique because it allows large data set to be represented by much smaller random sample of data

- Different methods of sampling are SRS, cluster sampling, systematic sampling, etc.

* Data Discretization -

→ • Data discretization means dividing the range of continuous attribute into intervals.

- It reduces number of values for a given continuous attribute

- It helps to a concise, easy-to-use, Knowledge-level representation of mining results

- Discretization techniques can be categorized as supervized & unsupervised discretization

- Supervised discretization uses class information & Unsupervised discretization does not used class information

- Top-Down :
  The process starts by first finding one or few points to split entire attribute range, & then repeats recursively on intervals. It is called Splitting

- Bottom-Up :
  It starts by considering all continuous values, removes some by merging neighbourhood values to form intervals, & repeats recursively. This is called Meging

★ Comparison -

	Data Science	Machine Learning	Artificial Intl.
i)	Data science is interdisciplinary field that uses scientific methods, processes to extract insights & knowledge from data	ML is subset of AI focused on development of algorithms & statistical models that enable computer to perform tasks without being programmed	AI is simulation of human intelligence processes by machines. It involves creation of algorithms that can learn, reason & solve problems
ii)	Techniques used are Data cleaning, Data analysis etc	Techniques used are Regression, Clustering etc	Techniques used are NLP, Robotics, etc.
iii)	Tools & Libraries required are Python, R, SQL, Excel	Tools & Libraries req. are TensorFlow, Keras, PyTorch, etc	Tools & Libraries req. are TensorFlow, NLTK, etc.
iv)	Application - Finance, Healthcare, Marketing, etc	Application - Image Recognition, Speech Recognition, etc	Application - Game Playing, Virtual Assistants

classmate
Date
Page

★ Feature Engineering -

→ • Feature Engineering is the process of cutting down the features.

• It plays a crucial role in enhancing model's ability to find patterns and make accurate predictions.

• Typically, feature engineering includes following aspects:

1) Handling Missing Data -
   • Addressing missing values by deciding whether to drop or fill them based on nature of data

2) Encoding Categorical Variables -
   • Converting categorical variables into format suitable for ML algorithms

3) Scaling & Normalization -
   • Scaling numerical features to ensure that they have similar magnitude.

4) Handling Outliers -
   • Identifying & Dealing with Outliers to prevent them from affecting performance of model.

* Regression -

→ • Regression is statistical method used in ML to model the relationship between a dependent variable & one or more independent variables

• There are several types of regression techniques,

1) Linear Regression -
   • It is the simplest & most widely used regression technique.
   • Equation for simple linear regression with one independent variable is:
     $y = mx + b$
   • Eg, Predicting House prices based on features like size, location, etc.

2) Polynomial Regression -
   • It is extension of linear regression where relationship between dependent & independent variable is modeled as 'n'th degree polynomial
   • Eg, modeling relationship between Temperature & Humidity

3) Ridge Regression -
   • It is regularization technique used to prevent overfitting in linear regression models by adding penalty term to loss function
   • Eg, Predicting student's GPA based on study hours, attendance, etc.

★ DSBDA ★

classmate
Date
Page

Statistical Interference

▲ UNIT - II ▲

Formulae :-

1) Mean - $\frac{\Sigma x_i}{n}$

2) Median - i) Arrange numbers in ascending order
ii) If odd numbers,
Median = middle value
iii) If even numbers,
Median = Average of middle two values

3) Mode - Value that occurs most

4) Variance -
(a) Sample : $s^2 = \frac{\Sigma (x_i - \bar{x})^2}{n-1}$

(b) Population : $\sigma^2 = \frac{\Sigma (x_i - \mu)^2}{N}$

5) Standard Deviation (S.D) -
(a) Sample : $\sqrt{s^2} = \sqrt{\frac{\Sigma (x - \bar{x})^2}{n-1}}$

(b) Population : $\sqrt{\sigma^2} = \sqrt{\frac{\Sigma (x - \mu)^2}{N}}$

Date
Page

★ Difference -

Standard Deviation	Variance
i) It is measure of dispersion of values of data set from their mean	i) It is statistical measure of how far numbers are spread in data set from their average
ii) It is common term in statistical theory to calculate central tendency	ii) It is primarily used for statistical probability distribution
iii) It measures absolute variability of dispersion	iii) It helps determine size of ~~spre~~ data spread
iv) Standard Deviation is symbolized by Greek Letter sigma 'σ'	iv) Variance is symbolized by 'σ²', i.e sigma squared
v) $\sigma = \sqrt{\frac{\Sigma(x-m)^2}{n}}$ M = mean, x = value in data n = no. of values	v) $\sigma^2 = \frac{\Sigma(x-m)^2}{n}$ m = Mean x = value in data n = no. of values
vi) Used in finance sector as measure of market	vi) Used in asset allocation

* Bayes Theorem -

→ • Bayes' Theorem is a method to revise the probability of event given additional information

• It calculates conditional probability called posterior or revised probability

• If A and B denote 2 events, P(A|B) denotes conditional probability of A occurring, given that B occurs

• Similarly P(B|A) denotes conditional probability of B occurring, given that A occurs

• Bayes' theorem gives a relation between P(A|B) and P(B|A)

• Eg, Suppose that $B_1, B_2, B_3 \dots B_n$ partition the outcomes of an experiment & that A is another event.
For any number k, with $1 \leq k \leq n$, we have

$$P(B_k|A) = \frac{P(A|B_k) \cdot P(B_k)}{\sum_{i=1}^{n} P(A|B_i) \cdot P(B_i)}$$

* Hypothesis Testing -

→ • A statistical hypothesis is a ~~test~~ procedure for deciding between two possible statement about a population.

• It is a method that uses sample data. to evaluate a hypothesis about a population.

• Goal is to analyze a sample in an attempt to distinguish between population characteristics that are likely and unlikely to occur.

■ Steps in Hypothesis Testing:

~~1) Specify~~

1) Formulate the Hypothesis -
   - Null Hypothesis is statement of status quo, one of no difference or effect.
   - Alternative Hypothesis is one in which some difference or effect is expected.

2) Select an appropriate test -
   - Test statistic measures how close sample has come to null hypothesis.

3) Choose Level of Significance -

▲ Type I Error :

- Occurs if null hypothesis is rejected when it is in fact true
- Probability of type 1 error (α) is also called level of significance

▲ Type II Error -

- Occurs if null hypothesis is not rejected when it is in fact false
- Probability of type II error is given by β
- Unlike α, magnitude of β depends on actual value of population parameter

• It is necessary to balance the two types of errors.

4) Collect Data and Calculate Test Statistic -

- Test statistic Z is calculated as,

$$Z_{cal} = \frac{\hat{p} - \pi}{\sigma_p}$$

5) Determine probability value / critical value -

- Determined using standard normal tables.

★ Wilcoxon Rank-sum Test -

→ • Wilcoxon Rank-sum test is non-parametric alternative to two sample t-test which is based solely on order of observations

• It is used to test null hypothesis that median of distribution is equal to some value.

■ Logic -

• Data are ranked to produce two rank totals, one for each ~~test~~ condition

• If there is systematic difference between two conditions, then most of high ranks will belong to one condition & most of low ranks to other

• As a result, one of rank totals will be quite small.

• On other hand, if two conditions are similar, then high & low ranks will be distributed fairly evenly betⁿ two conditions

• As a result, rank totals will be quite large.

• The Wilcoxon test statistic 'W' is simply the smaller of rank totals.

- Wilcoxon rank sum test rejects the hypothesis that two populations have identical distributions when rank sum W is far from its mean.

* Descriptive Statistics -

→ • Descriptive Statistics involves computing values which summarize set of data

- This typically includes statistics like mean, S.D, median, max, etc.

- Descriptive statistics is key because it allows us to present large amounts of raw data in meaningful way

- Descriptive statistics are also used to represent data graphically.

- Inferential Statistics -
  It is used to make inferences about the population.

* Chi-Square Test For Independence -

→ • The chi-Square test of independence is used to determine if there is significant relationship between 2 nominal variables.

• The frequency of each category for one nominal variable is compared across categories of second nominal variable.

• This test is also known as chi-Square Test of Association

• This test utilizes a contingency table to analyze data

• Data can be displayed in contingency table where each row represents category for one variable & each column represents category for other variable.

* Contingency Table -

→ • A contingency table is a statistical table used to display frequency distribution of two categorical ~~values~~ variables

• Contingency tables are useful for analyzing categorical data and identifying patterns or associations between variables.

• Eg, Suppose we have collected data on gender & smoking status of individuals in population

Then,
data is summarized in following contingency table:

	Non-Smoker	Smoker	Total
Male	200	100	300
Female	150	50	200
Total	350	150	500

• Contingency table are commonly used in Sociology, Psychology, Hypothesis testing, etc.

★ One-Tailed & Two-Tailed t-test :

→ ▪ One-Tailed t-test :

- If the hypothesis states that population parameter is greater / less than specific value. Then it is called one-tailed t-test.

- Eg,

Group A (Traditional method):

Mean score - 75 ; Standard Deviation - 10 ; Sample Size - 30

Group B (New method):

Mean score - 80 ; Standard Deviation - 12 ; Sample Size - 30

You can conduct one-tailed t-test to test if mean score of Group B is significantly greater than that of Group A

▪ Two-Tailed t-test :

- If hypothesis states that population parameter is different from specific value. Then it is called Two-Tailed t-test.

- Eg, Using the same data as for one-tailed t-test,

We can conduct a two-tailed t-test to test if there is any difference in mean scores of Group A & B.

Date
Page

* Population & Sample -

→ ▪ Population :

- Population refers to entire group of individuals, objects that you're interested in studying and drawing conclusions about.

- Eg, if we are studying heights of all males in particular country, Population would include heights of every single adult male in that country

▪ Sample :

- It is a subset of population that is selected for observation or analysis.

- The goal is to use characteristic of sample to make interferences about larger population.

NUMERICALS

Numerical on Consistency

Q] Find which driver is more consistent

Drive	Mark	Myra
1	28	27
2	22	27
3	21	28
4	26	6
5	18	27
	$\Sigma = 115$	$\Sigma = 115$

→

i) For mark, $\bar{x} = 23$

x	$x - \bar{x}$	$(x - \bar{x})^2$
28	5	25
22	-1	1
21	-2	4
26	3	9
18	-5	25
		$\Sigma = 65$

$$\therefore \boxed{S.D = \sqrt{\frac{\Sigma (x - \bar{x})^2}{N}}} = \sqrt{\frac{65}{5}} = \sqrt{13} = 3.605$$

$$\boxed{C.O.V = \frac{3.605}{23} \times 100} = 15.67\%$$

$$\frac{S.D}{\bar{x}} \times 100$$

ii) For Myra, $\bar{x} = 23$

x	$x - \bar{x}$	$(x - \bar{x})^2$
27	-4	16
27	-4	16
28	5	25
6	-17	289
27	4	16
		$\Sigma = 362$

$$S.D = \sqrt{\frac{\Sigma(x-\bar{x})^2}{N}} = \sqrt{\frac{362}{5}} = \sqrt{72.4} = 8.508$$

$$C.O.V = \frac{8.508}{23} \times 100 = 36.99\%$$

Hence, comparing both C.O.V,

Mark is more consistent, as he has lower C.O.V

- Numerical on Co-relation -

Q. Calculate Correlation between scores in two subjects.

R No	1	2	3	4	5	6	7
X	40	70	84	74	26	78	48
Y	64	74	100	60	50	48	80

→ Preparing Table,

X	Y	X^2	Y^2	XY
40	64	1600	4096	2560
70	74	4900	5476	5180
84	100	7056	10000	8400
74	60	5476	3600	4440
26	50	676	2500	1300
78	48	6084	2304	3744
48	80	2304	6400	3840
$\Sigma X = 420$	$\Sigma Y = 476$	$\Sigma X^2 = 28096$	$\Sigma Y^2 = 34376$	$\Sigma XY = 29464$

Here,

$$\bar{X} = \frac{\Sigma X}{N} = \frac{420}{7} = 60$$

$$\bar{Y} = \frac{\Sigma Y}{N} = \frac{476}{7} = 68$$

Correlation coefficient,

$$\gamma(X,Y) = \frac{cov(X,Y)}{\sigma_x \cdot \sigma_y}$$

$$= \frac{1/n \sum XY - \bar{X}\bar{Y}}{\sqrt{\left(\frac{1}{n}\sum X^2 - \bar{X}^2\right)\left(\frac{1}{n}\sum Y^2 - \bar{Y}^2\right)}}$$

$$= \frac{29464/7 - 60 \times 68}{\sqrt{\left(\frac{28096}{7} - 60^2\right)\left(\frac{34376}{7} - 68^2\right)}}$$

$$\gamma(X,Y) = 0.3749$$

- Numerical on IQR -

Q) Find IQR for,
[48, 52, 57, 64, 72, 76, 77, 81, 85, 88]

→ I) The given data set is in ascending order

II) Median = $\frac{72+76}{2} = \frac{148}{2} = 74$

III) Divide the data set in two parts.

∴ [48, 52, 57, 64, 72] & [76, 77, 81, 85, 88]

Let Q_1 = [48, 52, 57, 64, 72]

∴ Median = 57

Let Q_3 = [76, 77, 81, 85, 88]

∴ Median = 81

IV) IQR = $Q_3 - Q_1$

= 81 - 57

IQR = 24