Date 14/8/23 Page no:

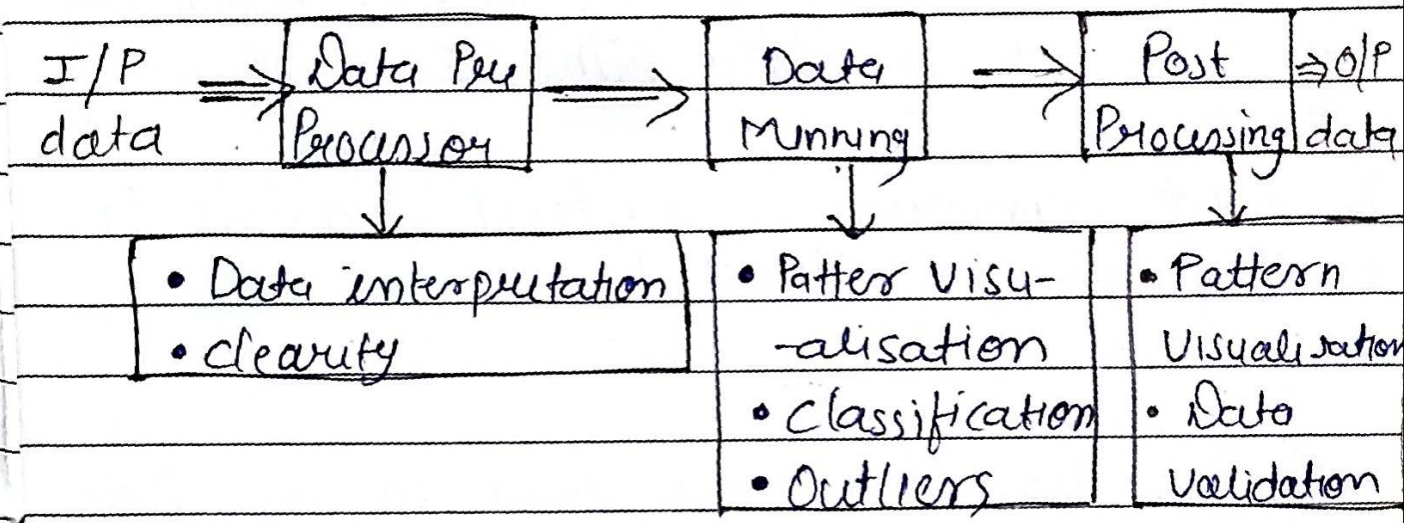
Data Science :-

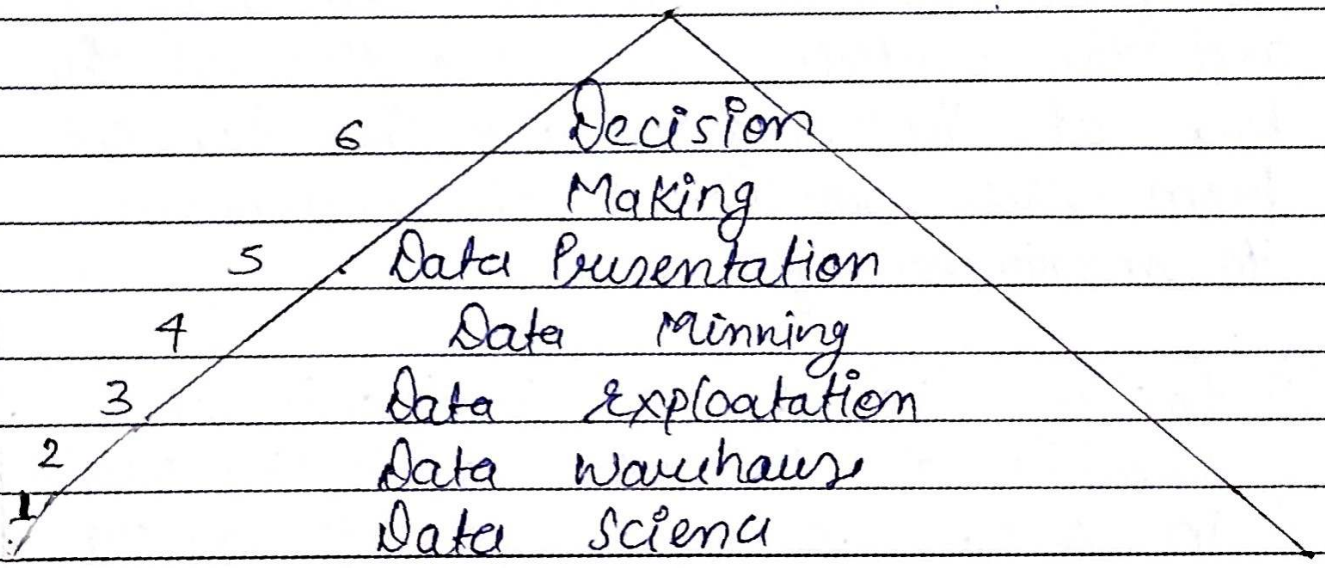
Abstracting information from a source is called data mining.

⇒ Compilation Process :-

1. Lexical analyser
2. Syntax analyser
3. Semantic
4. Intermediate
5. Code optimisation
6. Code generation

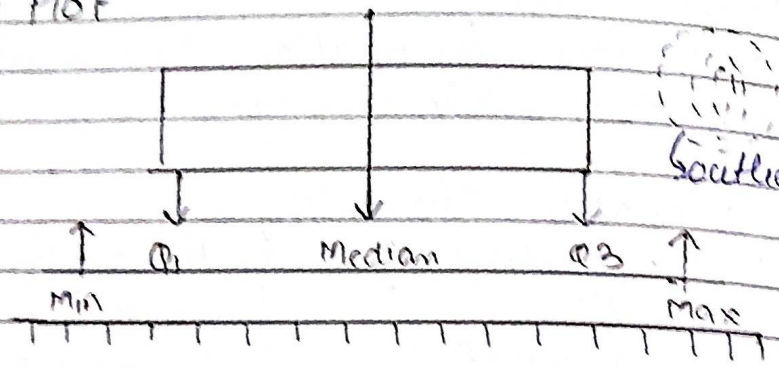
21/08/23





6 D'S

Box Plot



A box plot display the 5 no. summary of a set of data. The 5 no. summary is :-

1. first minimum
2. first Quartile Q_1
3. Minimum first Quartile
4.

In a box plot we draw a box from the first quartile to the 3rd quartile and the vertical line goes through the box at the median and the box go from each quartile to the minimum to maximum for ex :-

Q1) finding the 5 no. summaries. A sample of 10 boxes with the weight in gram i.e 25g, 28g, 29g, 30g, 34g, 35g, 37g, 38g, 29g, 35g make a box plot of the data

Step 1 :- Order the data from smallest to largest

25, 28, 29, 29, 30, 34, 35, 35, 37, 38

Step 2 :- find the median

$$\frac{30 + 34}{2} = 32$$

Step 3 :- finding 1st Quartile :-
The first quartile is the median of the data points to the left of the median i.e Q_1

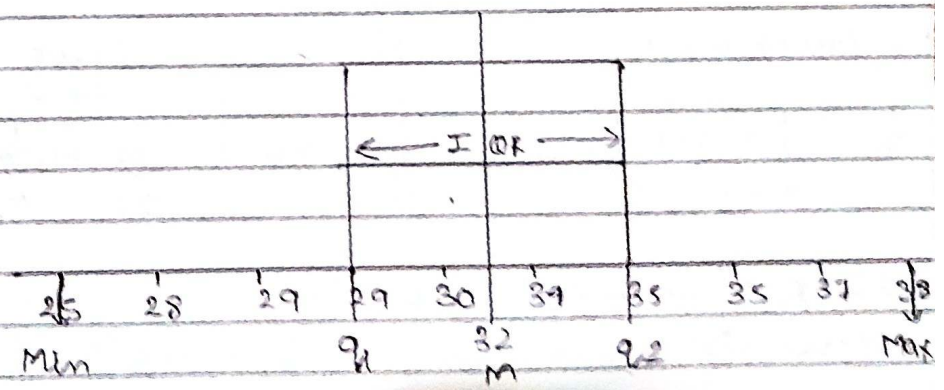
$\Rightarrow Q_1 = 29.$

Step 4 :- find 3rd Quartile :-
The third quartile is the median of the data points that is right of the median

$\Rightarrow Q_3 = 35$

Step 5 :- find Min and Max

$\Rightarrow$ Min = 25 , Max = 38



Date: / / Page no:

⇒ Inter Quartile range (IQR) = $Q_3 - Q_1$

⇒ IQR = 35 − 29 = 6

Q2) find out the 5 tuple summary with (IQR) let the data range be 199, 201, 236, 271, 278, 283, 291, 301, 303, 341 {∵ n = 11}

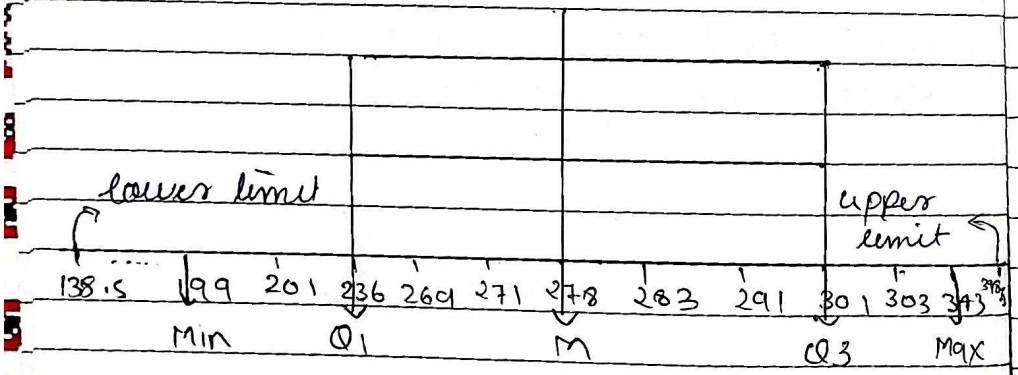
STEP 1 :- 199, 201, 236, 271, 278, 283, 291, 301, 303, 341,

STEP 2 :- M = 278

STEP 3 :- Q_1 = 236

STEP 4 :- Q_3 = 301

STEP 5 :- Min 199, Max 341



⇒ IQR = 301 − 236 = 65

⇒ lower limit = Q_1 − 1.5(IQR)

Date: / / Page no:

⇒ lower limit = 236 − 97.5

⇒ lower limit = 138.5

⇒ UPPer limit = Q_3 + 1.5 (IQR) = 301 + 97.5

⇒ upper limit = 398.5

List Square Method

Q1) A producer believes that the sale of his product for that years is related to an economic index and the data for that period is shown in the following tables

Years	Economic Index (x)	Sales (Y_n) 10000 Unit
1	104	2.0
2	100	2.3
3	111	2.1
4	129	2.6
5	126	2.3
6	115	2.4
7	152	2.5
8	161	2.8
9	143	2.6
10	170	3.0
n = 10	Σx = 1311	ΣY_n 24.6

(i) Determine the equation of the list Squares line that disrube the relation

b/w the sale and the economic indicator.

ii) Determine the strength of the relation b/w the two variably by computing the value of coefficient of ⓒ relation

iii) If the value of the economic index for 11th year is 175 find out the sales in the 11th year

Soln $Y = a + mx \Rightarrow Y_a = na + m\Sigma x$ — (1)

for for list square method

$\Sigma x Y_a = a\Sigma x + m\Sigma x^2$ — (2)

$\Rightarrow 24.6 = 10 \cdot a + m \cdot 1311$

Years	Economic Index (x)	Sales (Y_a) 10000 Units	$x \cdot Y_a$	x^2	y^2
1	104	2.0	208	10816	4.0
2	100	2.3	230	10000	5.2
3	111	2.1	233.1	12321	4.41
4	129	2.6	335.4	16641	6.76
5	126	2.3	289.8	15876	5.29
6	115	2.4	276	13225	5.76
7	152	2.5	380	23104	6.25
8	161	2.8	450.8	25921	7.84
9	143	2.6	371.8	20449	6.76
10	170	3.0	510	28900	9.0
	$\Sigma x = 1311$	$\Sigma Y_a = 24.6$	$\Sigma x Y_a = 32849$	177253	61.36

$$\rho_{(\gamma)} = \frac{n\Sigma xa - (\Sigma x)\cdot(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y_a^2 - \Sigma y^2}}$$

↳ coefficient of relation

$\Rightarrow 24.6 = 10a + 1311m$ — (3)

$\Rightarrow 3284.9 = 1311a + 177253m$ — (4)

multiply with 1311 on eqn (3) and 10 on eqn (4)

$\Rightarrow 32250.6 = 13110a + 1718721m$

$\Rightarrow 32849.0 = 13110a + 1772530m$

$+5984 = +53809m$

$\Rightarrow m = 0.011$

$\Rightarrow 24.6 = 10a + 14.421$

$\Rightarrow \boxed{a \simeq 1.01}$ ans//

1) $Y_a = \underset{a}{1.002} + \underset{m}{0.011}\,x$

2) $\gamma = \dfrac{n\Sigma xY_a - (\Sigma x)\cdot(\Sigma Y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma Y_a^2 - (\Sigma Y)^2}}$

$= \dfrac{10 \times 3284.9 = (1311)\cdot(24.6)}{\sqrt{10\times177253 - (1311)^2}\sqrt{10\times61.36 - (24.6)^2}}$

$= \dfrac{32849 - 32250.6}{\sqrt{53809}\times\sqrt{8.44}} = \dfrac{598.4}{673.84}$

$\Rightarrow \boxed{\gamma = 0.89}$ ans//

Date: / / Page no:

3) $Y_a = 1.002 + 0.011(x)$

$= 1.002 + 0.011(175)$

$\Rightarrow Y_a = 2.927$

Time Series Analysis

Ques 1) Find the Quarterly Sales of the fifth year by suitable forecasting method /technique for the data given 4 years also make adjustment for expected seasonal variation

Years	Quarter(x)	Sales 1000 Units	xYa	x^2	$Y' = 1.80 + 0.185x$
1	1	1.0	1	1	$Y_1 = 1.985$
	2	3.0	6	4	$Y_2 = 2.17$
	3	4.0	12	9	$Y_3 =$
	4	2.0	8	16	
2	5	1.0	5	25	
	6	3.0	18	36	
	7	5.0	35	49	
	8	3.0	24	64	
3	9	2.0	18	81	
	10	4.0	40	100	
	11	6.0	66	121	
	12	2.0	24	144	
4	13	2.0	26	169	
	14	5.0	70	196	
	15	7.0	105	225	
	16	4.0	64	256	
	$\Sigma x = 136$	$\Sigma Y_a = 54$	522	1496	

$n = 16$

$\Rightarrow$ We know that

$$\Sigma Y_a = na + m\Sigma x$$

$\Rightarrow$ putting values

$$54 = 16 \times a + m \times 136$$

$$54 = 16a + 136m \quad \text{——— (1)}$$

$\Rightarrow$ now,

$$\Sigma xY_a = a\Sigma x + m\Sigma x^2$$

$\Rightarrow$ putting values

$$522 = a \times 136 + m \times 1496$$

$$522 = 136a + 1496m \quad \text{——— (2)}$$

$\Rightarrow$ multiplying eqn (1) with 11 and then eqn (1) − eqn (2)

$$594 = 176a + 1496m$$

$$522 = 136a + 1496m$$

$$72 = 40a$$

$$\Rightarrow a = \frac{72}{40} = 1.80$$

$\Rightarrow$ put a in eqn (1)

$$54 = 16 \times 1.80 + 136m$$

$$54 = 28.8 + 136m$$

$$54 - 28.8 = 136m$$

$$25.2 = 136m$$

$$\Rightarrow m = \frac{25.2}{136} \Rightarrow 0.185$$

So $y' = 1.80 + 0.185x$

now for 5th year, we have to find quarterly sales of 5th year i.e for $x = 17, 18, 19, 20$ so

$\Rightarrow y^{17} = 1.80 + 0.18(17) = 4.945$

$\Rightarrow y^{18} = 1.80 + 0.18(18) = 5.13$

$\Rightarrow y^{19} = 1.80 + 0.18(19) = 5.315$

$y^{20} = 1.80 + 0.18(20 = 5.5$

Ques2) A machine shop produces steel pins the width of the 100 pin was checked after manufacturing and data was recorded as follows

width (in mm)	freq (f)	fx_i	x^2	fx_i^2	Cell mid point (x_i)
9.50 - 9.51	6	57.03	90.34	542.04	9.505
9.52 - 9.53	2	19.05	90.72	181.44	9.525
9.54 - 9.55	20	190.99	91.10	1822	9.545
9.56 - 9.57	32	306.08	91.48	2927.36	9.565
9.58 - 9.59	22	210.87	91.87	2021.14	9.585
9.60 - 9.61	8	76.14	92.25	738	9.605
9.62 - 9.63	6	57.75	92.64	555.84	9.625
9.64 - 9.65	4	38.58	93.02	372.08	9.645
	$\Sigma f_i = 100$	957.1	733.42	9159.9	

↳ find out arithmetic mean, standard deviation and variance (H.W)

Formula :- 1) Mean :- $\bar{x} = \frac{\Sigma f_i x_i}{\Sigma f_i}$

2) Standard deviation :- $\sigma = \sqrt{\frac{\Sigma f x^2}{n} - (\bar{x})^2}$

3) Variance $= \sigma^2$ 28|08|23

Sol^n ① Mean $= \bar{x} = \frac{\Sigma f_i x_i}{\Sigma f_i} = \frac{957.1}{100} = 9.571$

$\Rightarrow$ Mean $= 9.571$

② Standard deviation :- $\sigma = \sqrt{\frac{\Sigma f x^2}{n} - (\bar{x})^2}$

$\Rightarrow \sigma = \sqrt{\frac{9160.12}{100} - (9.572)^2}$

$\Rightarrow \sigma = \sqrt{91.601 - 91.604}$

$\Rightarrow \sigma = \sqrt{-0.03} \Rightarrow \sigma = 0.03i$

③ variance $= \sigma^2 = 0.03$

Ques3)

The measure values are 344°, 338°, 342° 335° and 336° their values are constitute the first sub group

Date: / / Page no:

Frequency Distribution:-

Frequency distribution is a tabulation of data obtains from measurement arranged in ascending or descending order acc. to size and find out ~~determine the~~ range of the data, arithmatic mean, mid point, median & mode

Date: / / Page no:

$$\Rightarrow \sigma = \sqrt{\frac{(335-339)^2 + (336-339)^2 + (338-339)^2 + (342-339)^2 + (344-339)^2}{5}}$$

$$\Rightarrow \sigma = \sqrt{\frac{16+9+1+9+25}{5}} = \sqrt{\frac{60}{5}}$$

$$\Rightarrow \sigma = \sqrt{12} \Rightarrow \boxed{\sigma = 3.46} \text{ Std deviation}$$

$$\Rightarrow \boxed{\sigma^2 = 12} \rightarrow \text{Variance}$$

Ques 3) 5 theoritical Control are listed to determine the temp the measure are 344, 338, 342, 335 and 336 their values are constitute the 1st subgroup compud Arithmetic mean, median, range, variance, Std. deviation

Soln arranging it in ascending order

$\Rightarrow$ ~~344~~, 335, 336, 338, 342, 344

$\Rightarrow$ Median = 338

$$\Rightarrow AM(\bar{x}) = \frac{335+336+338+342+344}{5} = 339$$

$\Rightarrow$ Range = 344 − 335 = 9//

$$\Rightarrow \text{std deviation} = \sqrt{\frac{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \ldots (x_n-\bar{x})^2}{5}}$$

Ques 4) The mean and the standard deviation of a sample of 100 observation was calculated as ~~40~~ and 5.1 respectively while company with the original data it was found that by mistake a figure of 40 was miscopied as 50 for one observation calculate the corrected mean and std. deviation.

Soln $\because AM(\bar{x}) = \frac{\Sigma X}{N}$

$\because \sigma = \sqrt{\frac{\Sigma x^2}{N} - (\bar{x})^2}$

(1) $\bar{x} = \frac{\Sigma X}{N} \Rightarrow 40 = \frac{\Sigma X}{100}$

$\Rightarrow \Sigma X = 4000$

$\Rightarrow \Sigma X = 4000 - 50 + 40 = 3990$

$\Rightarrow \boxed{\Sigma X = 3990}$

$\bar{x} = 39.90$

$\Rightarrow$ Std deviation (σ) $= \sqrt{\frac{\Sigma x^2}{N} - (\bar{x})^2}$

$\Rightarrow 5.1 = \sqrt{\frac{\Sigma x^2}{100} - (40)^2}$

$\Rightarrow (5.1)^2 = \frac{\Sigma x^2}{100} - 1600$

$\Rightarrow 26.01 = \frac{\Sigma x^2}{100} - 1600$

$\Rightarrow (26.01 + 1600) \times 100 = \Sigma x^2$

$\Rightarrow \Sigma x^2 = 1626.01 \times 100$

$\Rightarrow$ (a.m.c) $\Sigma x^2 = 162601 - 50^2 + 40^2$

$162601 - 2500 + 1600$

$\Sigma x^2 = 162601 - 900$

$\Sigma x^2 = 161701$ → correct

Correct Std. dw. $= \sqrt{\frac{\Sigma x^2}{100} - (\bar{x})^2}$

$= \sqrt{\frac{161701}{100} - (39.90)^2}$

$= \sqrt{\frac{161701}{100} - 1592.01}$

$\Rightarrow \sqrt{1617.01 - 1592.0}$

$\Rightarrow \sqrt{25}$

$\sigma \Rightarrow 5$ // → correct

Date 29/08/23 Page no.

Hypothesis Testing

↳ State H_0 as well as H_a

↳ Specify the value of significant (or the α value)

↳ Decide the correct Sampling Distribution

↳ Sample a random sample & work out an appropiate value for sample data

↳ Calculate the probability if H_0 is true

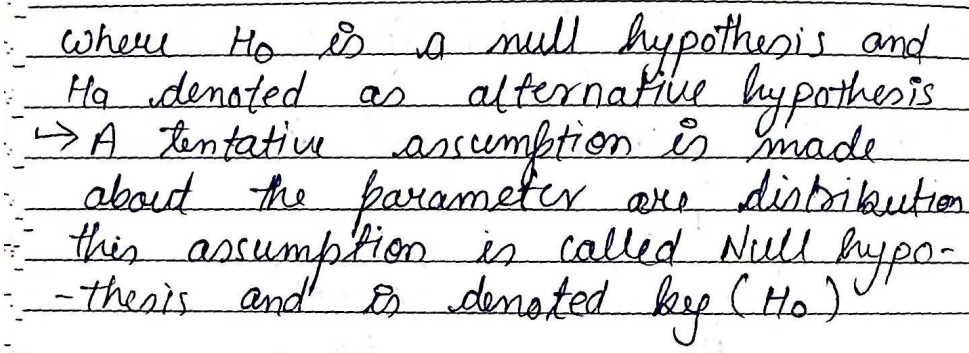
↳ Is this probability equal to or similar

↓ Yes	↓ NO
Reject H_0	Accept H_0
↓	
run this risk of committing Type I error	run this risk of committing Type II error

Date: / / Page no:

Flow diagram of Hypothesis testing

To test a hypothesis testing is to tell on the basis of the data the researchers are collected whether or not the hypothesis seems to be valid In hypothesis testing the main question is whether to accept the null hypothesis or not to accept the null hypothesis. Hypothesis testing can also be defined in the following flow diagram for better understanding



where H_0 is a null hypothesis and H_a denoted as alternative hypothesis

↳ A tentative assumption is made about the parameter are distribution this assumption is called Null hypothesis and is denoted by (H_0)

↳ An alternative hypothesis which the opposite of null hypothesis. In alternative hypothesis maintain the relation b/w two measured variables. The hypothesis testing procedure involves using sampling data to determine whether or not H_0 can be rejected. The ideal procedure is in hypothesis testing leads to acceptance of H_0 when H_0 is true, and the rejection of H_0 when H_0 is false. The hypothesis testing is based on the sampling information and possibilities of error must be considered

- A Type I error corresponds to rejecting H_0, where H_0 is actually true
- And a Type II error corresponds to accepting H_0, where H_0 is false.

The probability of making a Type I error is denoted by α, and a probability of making a Type II error is denoted by β

★ Z test :-

$$Z = \frac{\bar{x} - \mu_{H_0}}{\sigma_p / \sqrt{n}}$$

↳ Population normal infinite

- Sample size may be large or small but variance of the population is known
- H_a may be one sided or two sided
in such cases Z test is used

↳ Population normal finit
- Sample size may be large or small but variance of the population is known
- Ha may be one sided or two sided

in such case the z test is called finite population multiplyer

$$Z = \frac{\bar{X} - \mu H_0}{(\sigma_p/\sqrt{n}) \cdot \sqrt{(N-n)/(N-1)}}$$

↳ Population normal infinit
- Sample size small, variance of the population unknown
- Ha maybe one sided or two sided.

in such cases T test is used

$$t = \frac{\bar{X} - \mu H_0}{\sigma_s \sqrt{n}}$$

↳ Population normal finite
- sample size small, variance of the population unknown
- Ha may be one sided or two sided

in such cases t test is used

$$t = \frac{\bar{X} - \mu H_0}{(\sigma_s/\sqrt{n})\sqrt{(N-n)/(N-1)}}$$

↳ Population may not be normal but
- Same size, large variance of the
- population may be unknown or known and Ha maybe one sided or two side

in such situation, z test is used. (we can use either case 1 or case 2 formula)

for variance :- $\sigma_s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{(n-1)}}$

(Ques 1) A sample of 400 male students is found to have a mean height is 67.47 inches can it be reasonably regarded as a sample from a large population with mean height 67.39 inches and standard deviation is 1.30 inches test at 5% level of significance

Sol" $H_0 = \mu H_0 = 67.39"$

$H_a = \mu H_a \neq 67.39"$

$\bar{X} = 67.47"$, $\sigma_p = 1.30$, $n = 400$

$$Z = \frac{67.97 - 67.39}{1.31/\sqrt{400}} = \frac{0.08}{0.065}$$

$$\boxed{Z = 1.23}$$

Considering Ho is two sided for given question

Ques

Ques) The specimen of a copper wire drawn from a lodge set and have the following breaking strength in (weight in kg) 570, 572 578, 568, 572, 578, 570, 572, 596, 554 test using student whether the breaking strength of any lot may be taken to be 578 kg wait 5% level of significance. Verify degree of freedom, arithmetic mean, standard deviation and t test result

Date: 4/9/23 Page no:

#=Data warehousing :- Data warehousing governing all the data and control multidimensional opau data. The construction of data warehouse involves

- Data cleaning
- Data integration
- Data transformation

Data warehouse provides online analytical processing tool (OLAP) for interactive analysis of multidimensional data which facility effective data generalisation and data mining. Many other data mining functions such as association, classification, prediction and clustering can be integrated with OLAP operation to enhance interactive mining of knowledge in multiple level of abstraction.

Data warehousing providing architecture and tool for business execution to systematically organising understand and use their data to make strategic decision. It is a subject oriented integrated, and time variant and non volatile collection of data in support and management decision marketing process

Date: / / Page no:

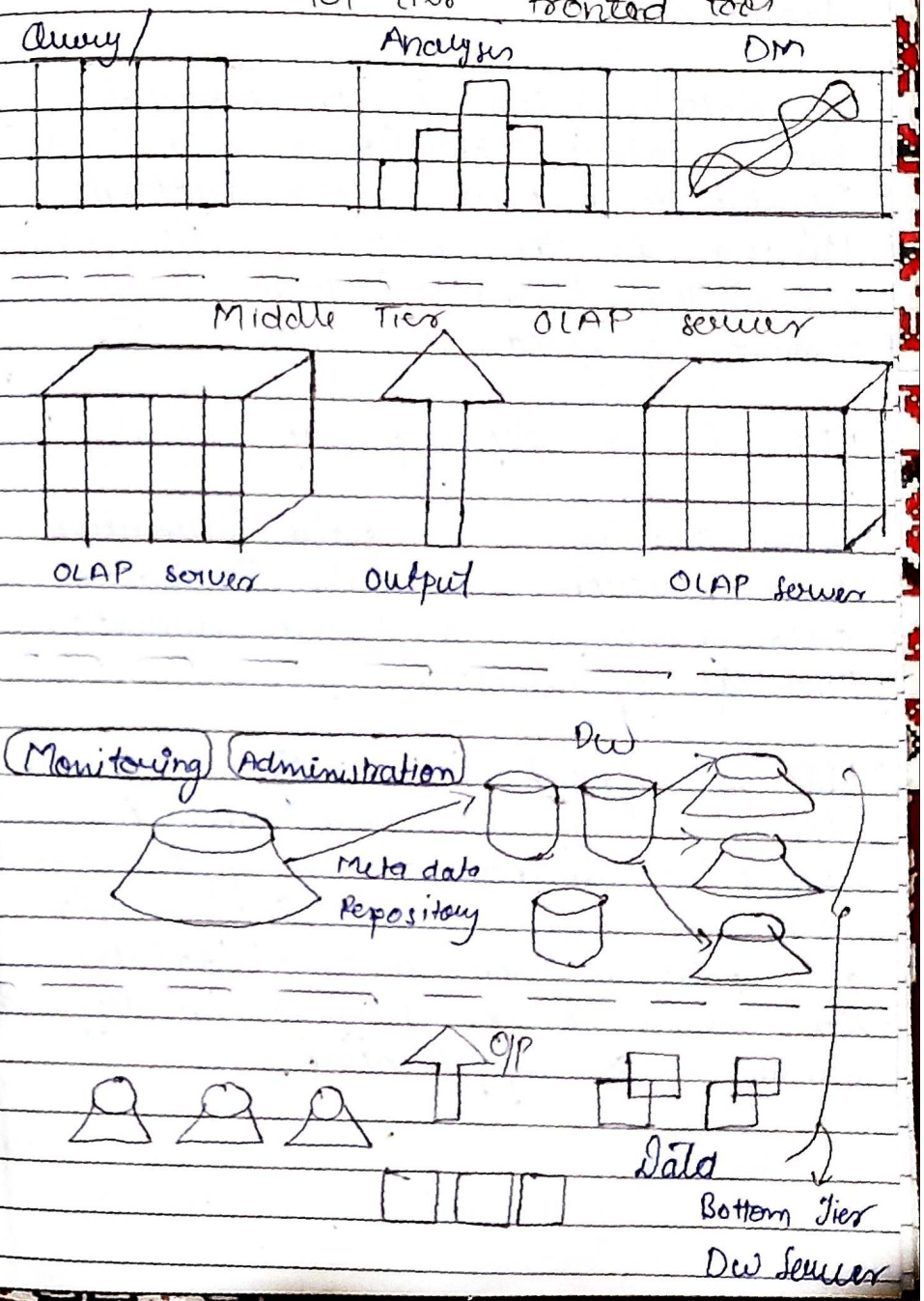
Compare OLAP & OLTP processing

The major distinguish feature of OLAP and OLTP are summarised as follows

features	OLTP	OLAP
User and System oriented	Customer oriented and used for transaction & query processing	Market oriented and used for data analysis
Data contained	Manage current data and easily used for decision making	Large amount of data manage historical data summarisation aggregation facility
Database design	E-R diagram data model, appl" oriented and DP design	star & snow fall model and subject oriented database design
Access pattern	Atomic transaction, concurrency control	Read only operations and complex queries
Users	Customer Oriented	Analyst, manager, it is market oriented

Date: / / Page no:

Multitier Architecture



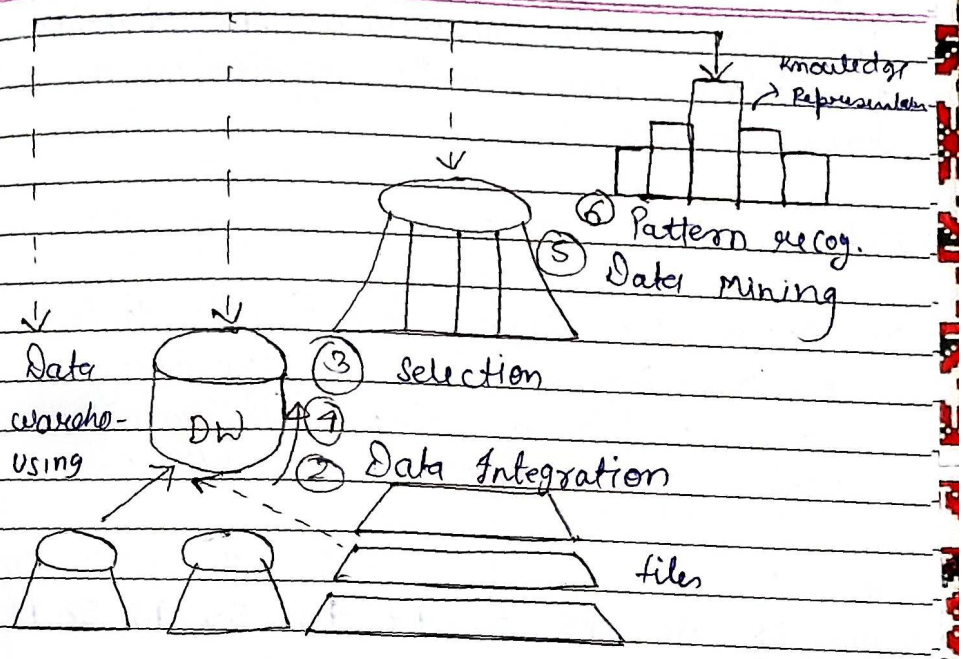
KDD Process (Knowledge Discovery from Data)

A data mining is an essential process where intelligence method are applied to extract data patterns. Mining is the process of discovering interesting knowledge from large amount of data.

KDD is a step by step processing

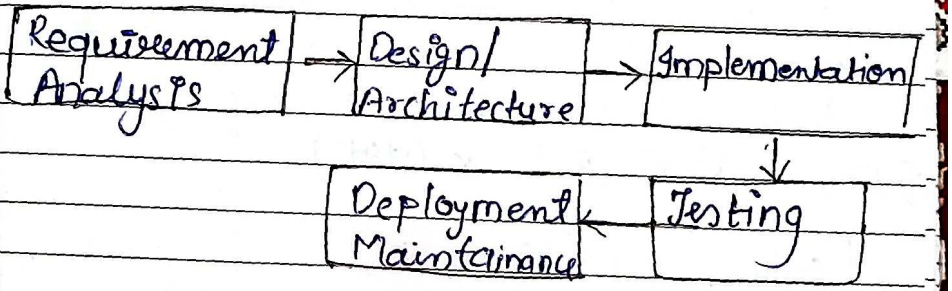
1. Data cleaning
2. Data Integration
3. Data selection
4. Data Transaction
5. Data mining
6. Pattern evaluation
7. Knowledge Representation

KDD is refered to as Knowledge Discovery in database and is defined as method of finding, transforming, and refining meaningful data and patterns from a raw database in order to be utilised in different domains or application.



05/09/2023

⇒ 5 phases of Software development:



SDLC

★ Internal 1st topic :- KDD process, Numerical on time series, Numerical on least square method, Theory on

box plot with Numerical, Introduction of data science with application of data science

Chi-Square Test :-

(Ques) Apply chi-square test in the context of sampling analysis for comparing a variance to a theoretical variance weight of 10 students as follows:-

S.no	1	2	3	4	5	6	7	8	9	10
Weight	38	40	45	53	47	43	55	48	52	49

Can we say that the variance of the distribution of weight for all students from which above sample of 10 students was = 20 Kgs. Test ther at 5% level of significance

Soln $x^2 = \frac{\sigma_s^2}{\sigma_p^2} \times (dof)$

S.NO.	Weight (x_i) in Kg	$(x_i - \bar{x})$
1	38	-9
2	40	-7
3	45	-2
4	53	6
5	47	0
6	43	-4
7	55	8
8	48	1
9	52	5
10	49	2

$\Sigma x_i = 470$

$\bar{x} = \frac{\Sigma x_i}{n} = \frac{470}{10} = 47$

$\Rightarrow \sigma_s = \sqrt{\frac{(x_i - \bar{x})^2}{n-1}}$

$\Rightarrow \sigma_s = \sqrt{\frac{(-9)^2+(-7)^2+(-2)^2+(6)^2+0^2+(-4)^2+(8)^2+(1)^2+(5)^2+(2)^2}{9}}$

$\Rightarrow \sigma_s = \sqrt{\frac{280}{9}} \Rightarrow \sigma_s = \sqrt{31.11}$

Now,

σ_s^2 = variance = 31.11

Now,

$x^2 = \frac{31.11}{20} \times 9 = 13.99$

dof = 9

at 5% $x^2 = 16.928$

Null hypothesis accepted