23/2/24

11

MARCH'23

SATURDAY

10th Week • 070-295

stats - 1

MARCH 2023

SUN	MON	TUE	WED	THU	FRI	SAT
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

09.00

stats ⇒ part of the math.

10.00

defn ⇒ collect, organize, analyze

11.00

12.00

Data ⇒ facts or piece of informatn.

Eg - IQ of students.

13.00

- Height of students in classroom.

14.00

- weight, age of the people.

15.00

16.00

2 types of stats

17.00

descriptive stats

inferential stats

18.00

def :- Summarizing our data with the single value.

one value

def :- find out the conclusion for based on sample data.

city → sample

population

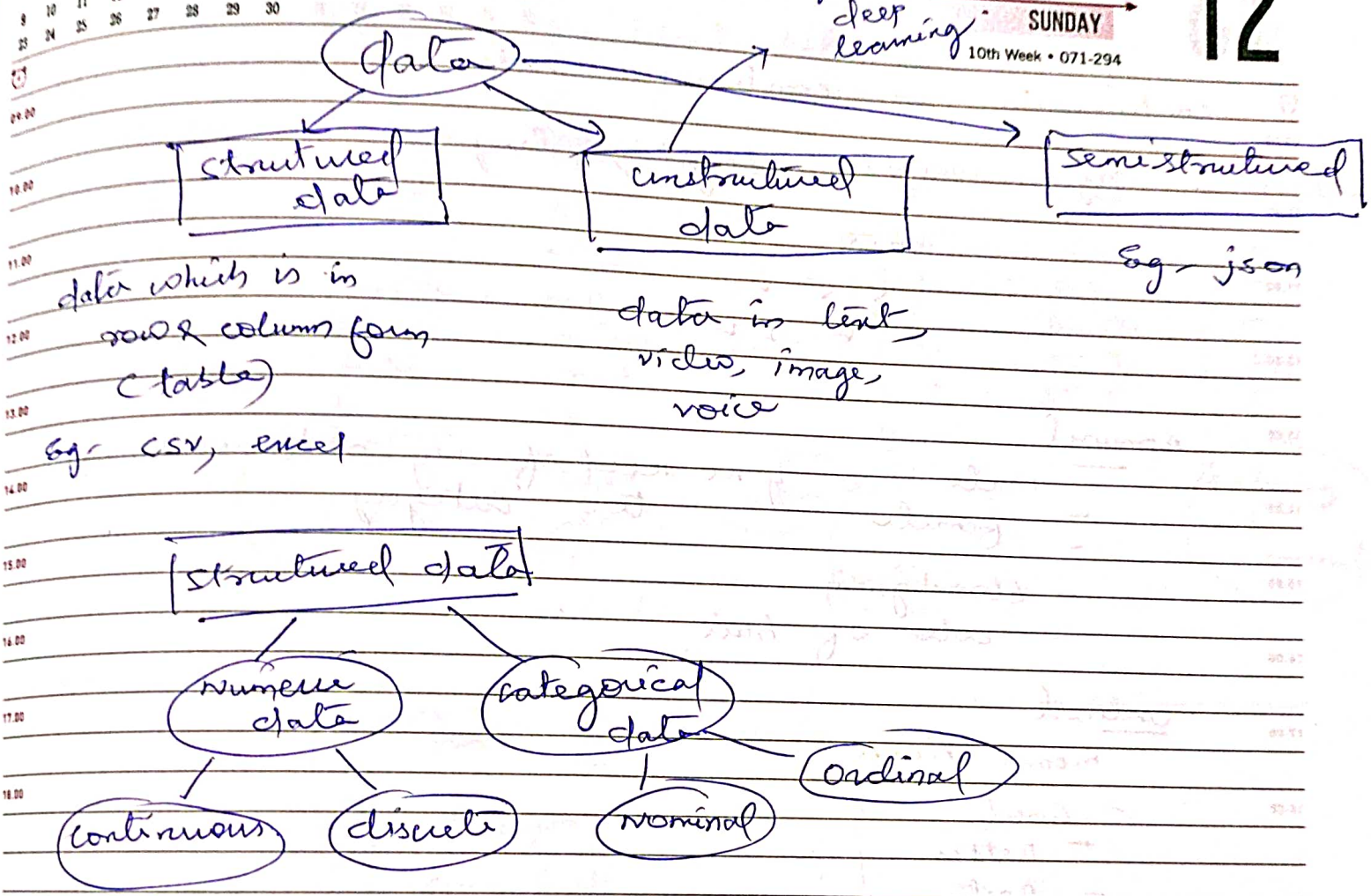
hypothesis

- CI test, p, one tail, two tail test, standard error, sig test.

Test

Z-test, t-test, anova, chi-square.

Nothing is impossible, Even the word impossible says i-m-possible.



data tendency
~
whether our data is best data or not we can find in data engineering. by these methods
- batch data
- mini batch data
- streaming data

discrete = whole no values.
Eg- no of students in class ~
60.5 X
it can be -ve.

SUN	MON	TUE	WED	THU	FRI	SAT
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

continuous - continuous value.

Eg - weight, height, temp

~~5.0~~

5.5 ✓

6.1 ✓

nominal

- male
- female

} no need of any order with respect to category.

- blood group
- color of hair

ordinal

means Rank.

- Good
- Better
- Best

↑

Eg: → Education ↑ 10, 12, BE, BTECH, phd ↓ they have rank.

→ Rollno

* Each column is called attribute / feature.

Given a structured data.

Eg!-

Rollno	Name	Degree	%	Height	weight	Gender
1	Amit	BE	85	5.5	60	m
2	Sunny	Btech	80	5.3	65	m
3	Suresh	Phd	75	6.1	62.2	m
4	komal	mtech	82.8	5.2	50.2	F
5	Ashi	MLA	95.5	5.7	55.5	F

- Name ↓ category ↓ nominal
- Degree ⇓ category data ↓ ordinal — order matters
- % ↓ numeric ↓ continuous
- Height ⇓ num ↓ continuous
- Gender: category ↓ nominal

Eg) semi-structured data:

name	description	sentiment
sunny	sunny is ds who is in memory good experience is m	positive

⇑ text data requires NLP approach; text preprocessing

descriptive stats
- ↳ central tendency of the data.
- ↳ dispersion method
- ↳ probability theory
- ↳ probability distribution func.

15 MARCH'23
WEDNESDAY
11th Week • 074-291

MARCH 2023

SUN	MON	TUE	WED	THU	FRI	SAT
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

09.00
10.00
11.00
12.00
13.00
14.00
15.00
16.00
17.00
18.00

inferential stats
- sample / populatn
- hypothesis testing
- z-test
- T-test
- F-Test
- anova test
- chi-square test

Descriptive stats:

it will give single o/p.

1. mean
2. median
3. mode

mean ⇒ {1, 1, 2, 2, 3, 3, 4, 5, 5, 6}

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

$$= \frac{1+1+2+2+3+3+4+5+5+6}{10}$$

$$= 3.2 \quad \text{(single value)}$$

population } inferential
sample } stats

mean → {10, 20, 30, 40, 50, 60, 70} ⇒ {10, 40, 50, 60}
(population) (sample)

Economy is a saving bank, into which men drop pennies, and get dollars in return.

in sample populatn we have to choose a sample data from populatn data.

measure	Populatn parameter	Sample statistic
mean	μ	$\bar{x}$
variance	σ^2	s^2
standard deviatn	σ	s

eg our populatn data = {10, 20, 30, 40, 50, 60, 70, 80, 90, 100}

mean → μ

standard deviatn → σ

variance → σ^2

from populatn data we have to choose sample data.

so {10, 50, 60, 90}.

mean → $\bar{x}$

standard deviatn → s

variance → s^2

Median → First sort the value then find middle value.

eg - 1, 2, [3], 4, 5

- 1, 2, [3, 4], 5, 6

$= \frac{3+4}{2} = 3.5$

1, 2, 3, 4, (10) ← this is a outlier ; it will create problem as it is not in sequence and finding its average value will create problem.

1, 2, 3, 4, (50) 60, 70, 80, 90, 100

↑ balancer

mode :- most frequent value in the dataset

(1) central tendency — mean, median, mode.

(2) Dispersion of data — Range, %, percentile, quartile (5 no Summary of the data), variance, standard deviation, covariance, correln.

data kitna fela hua hai.

Covariance - relⁿship bet two variable.

(3) probability theorem, random variable, probability distribution func.