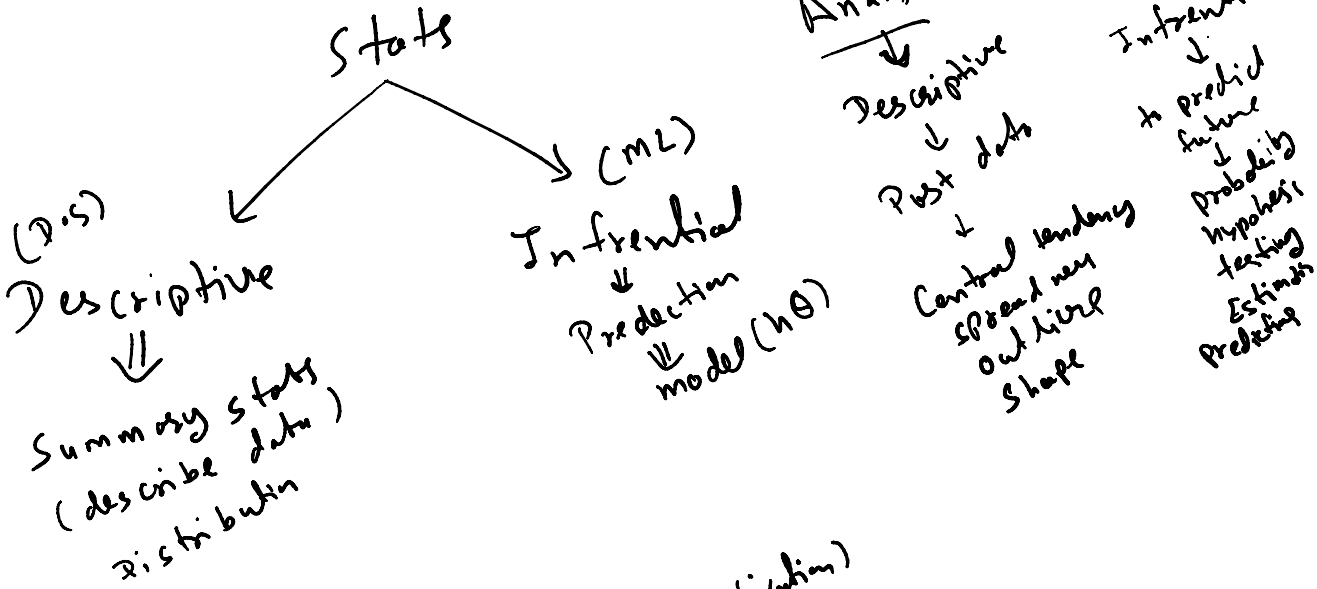
Day-1_Stats

Thursday, June 4, 2020 12:43 PM

Study of data to find some insights



Analysis vs Analytics

Descriptive → Past data → Central tendency, spread, outliers, shape

Inferential → to predict future → probability, hypothesis testing, Estimate, Prediction

⇒ Descriptive Stats →

Collection, Organisation (classification)

Summerization (Analysis)

presentation (Visulation)

→ Report explain data and its Behaviour

⇒ Inferential stats

→ Generalization from sample to population

→ hypothesis testing

→ model creation (prediction)

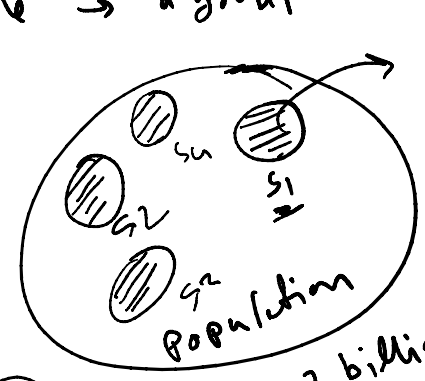
→ Estimation

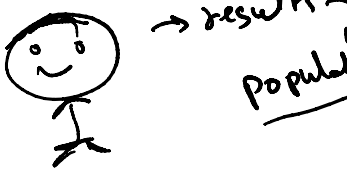
Report → talks about future estimation and predictions

. Live Data Set

Population → Entire Data Set

Sample → a group or example from data

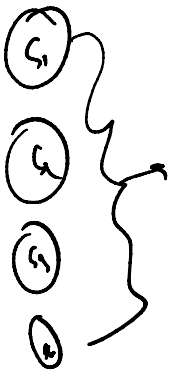




→ is india a census country ?.

Error prone ?.

No ?.



→ Inferential stats
hypothesis tests
↓
population

(1) Data type

(ii) Quantative Data (Numerical)

(iii) Qualitative Data (Categorical)

Quiz :

for each variable bielow, identify each as either quantitative or categorical

- Zip Code → Categorical → Zipcode
- Age → quantative
- Income → quantative
- Marital Status (Single, Married, Divorced etc.) → Categorical
- Height → quantitative

304504
2021
2023

} numerical data

Quantitative ?.

1. Age → quantitative
2. Income → quantitative
3. Marital Status (Single, Married, Divorced etc.) → Categorical
4. Height → quantitative
5. Letter Grades (A+, A, A-, B+, B, B-,...) → Categorical
6. Travel Distance to Work → quantitative
7. Ratings on a Survey (Poor, OK, Great) → categorical
8. Temperature → quantitative
9. Average Speed → quantitative
10. categorical

Categorical ? group

Age group

child — 0 - 16
young — 16 - 30
middle — 30 - 50
old — 50 - 70
versoud 70 above

304 50
30 2021
30 2023
30 30 21
30 30 70
50 20 21
30

926034
343748

Quantitive ?
mathematical function
summation zipcode ?
mean of zipcode ?
min & max ?

Age → 10, 20, 30, 25, 40, 47

avg age ?
min age ?
max age ?
} Quantative

Ordinal vs nominal

A
A+
B
B-
} order → ordinal

O 1. Letter Grades (A,B+,B,B-)
N 2. Types of Fruit(Apple,Banana,etc.) → we can not order them
O 3. Ratings on a Survey(Poor,Ok,Great)
N 4. Types of Dog Breeds(German Shepherd,Collie,etc.)
N 5. Genres of Moivies(Horror,Comedy,etc.)
N 6. Gender
N 7. Nationality
O 8. Education(HS,Associates,Bachelors,Masters,PhD,etc.)

Great
ok
poor

School
HS
BE
Master
PhD

Continous vs Descrete

- Continous — Travel Distance from Home to Work → km, meters, centi, milli ....
- Descrete — Number of Pages in Book → count
- Continous — Amount of Rain in Year → cm, milli, inches
- Continous — Time to Run a Mile → minute, sec, min
- Descrete — Number of Movies Watched in a Week → countable
- Continous — Amount of Water Consumed in a Day → ml, l,
- Descrete — Number of Phones per Household → countable

160, 165, 140, 137
[100, 150, 110, 105, ...]

145

Quantative Data Analysis

1. Measure of Center → single value that can describe properly of your entire data set or center point of data which describe majority points in data.

data

① Measure of Center → ... center, majority po...

Central tendency

② Measure of spread → Represents variability in data, generalize your central tendency

③ Shape of Data → Distribution of data, property of data

④ Outlier → extreme high or extreme low values

Assignment

find Data Distributions?.

normal
uniform
gausian
Gaussian
pdf

10 → scaling, rojat, 6 - Job

X 10,000 + 20,000

8000 - 12000

← 10

Accuracy ↓ X

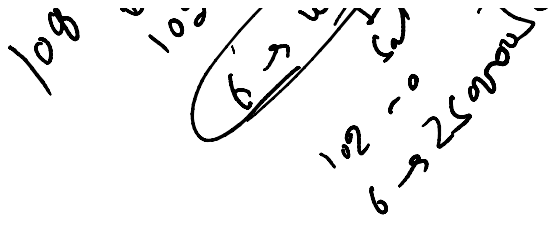
incorrect → outliers

10 → 10 - student (majority)

6 → working, sal ... 20000

2000

10



(1) Numpy — Vectors & Metrices

↓

Vector & matrices (Array)

Data Can be stored in →

1. Scalers
2. Vectors
3. matrix
4. Tensors

① Scalar is a single discreate observation

only magnitude

5

10

35.5

Each discreate observation is a scalar type

② row vector or Column Vectors

(2) row vector or col

1D Array

$\vec{v}$

Vector → collection of scalars

$$V_1 = \begin{bmatrix} 10 \\ 20 \\ 30 \end{bmatrix}$$ column vector

$$V_2 = \begin{bmatrix} 50 & 100 & 200 \end{bmatrix}$$ row vector

(3) Matrix → collection of vector

	qnt	price
p1 →	3	2000
p2 →	5	1500
p3 →	6	300

Shape → row x coll ⇒ (3, 2)

Size = row * cols ⇒ 6

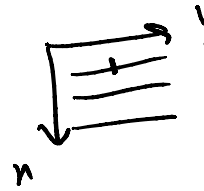
row → records, tuple

cols → features, attributes

Dimension of matrix - 2D

→ y

Dimension of mat.

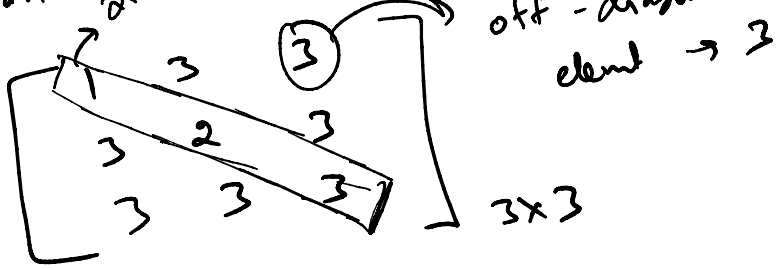


3 x 2
1 1

Square matrix

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} \quad 3 \times 3$$

Diagonal diagonal elements [1 2 3]



d = [1 2 3]

od = 3

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

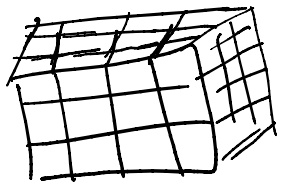
eye, Identity matrix

3 x 3

$$\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}_{3\times 2}$$ zero matrix

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}_{2\times 3}$$ ones matrix

→ Tensors → high dimensional arrays



4D - 5D

→ Operations on a matrix

① Transpose

$$A = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}_{3\times 2} \qquad A^T = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}_{2\times 3}$$

② Symmetric matrices

② Symmetric matrices

if $A = A^T$ then A is symmetric

$$A = \begin{bmatrix} 3 & 4 \\ 4 & 6 \end{bmatrix}_{2\times 2}$$

$$A^T = \begin{bmatrix} 3 & 4 \\ 4 & 6 \end{bmatrix}_{2\times 2}$$

$$\boxed{A = A^T}$$

① Addition

· element wise add

·· Cond: shape of Both matrix should be same

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \quad B = \begin{bmatrix} 7 & 8 & 9 \\ 10 & 11 & 12 \end{bmatrix}$$

$$A + B = \begin{bmatrix} 8 & 10 & 12 \\ \cdot\cdot & \cdot\cdot & 18 \end{bmatrix}$$

$$A+B = \begin{bmatrix} 8 & \cdots & \\ 14 & 16 & 18 \end{bmatrix}$$

$$A = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \end{bmatrix} \quad B \begin{bmatrix} B_{11} & B_{12} & B_{13} \\ B_{21} & B_{22} & B_{23} \end{bmatrix}$$

$A+B$, commutative

$$A+B = B+A$$

Stats

numpy