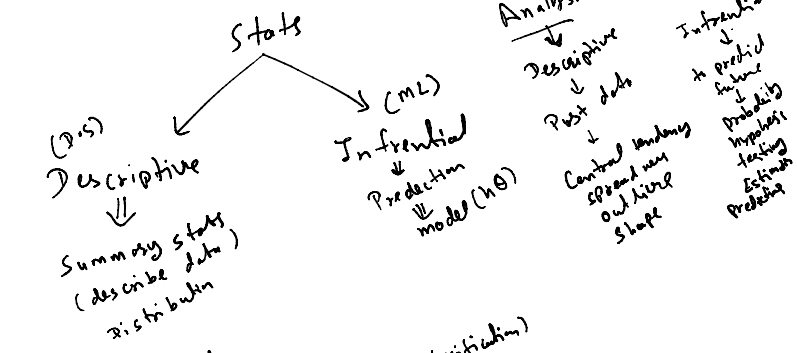
Day-1_Stats

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)
summerisation (Analysis)
presentation (Visualisation)

→ Report explained data and its Behaviour

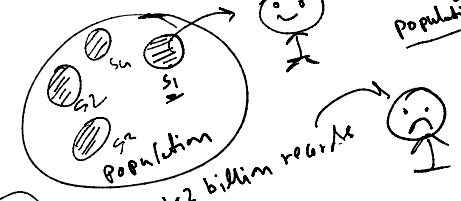
⇒ Inferential stats

→ Generalization from sample to population
→ hypothesis testing
→ model creation (prediction)
→ Estimations

Report → talks about future estimation and predictions

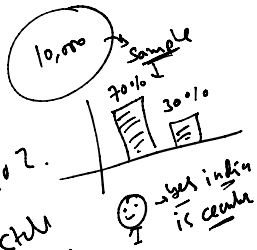
Population → Entire Data Set

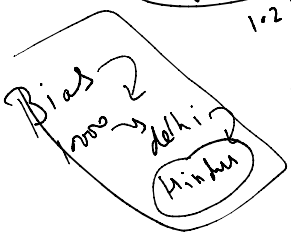
Sample → a group or example from data



1.2 billion records

→ is india a cercle country ?.





Error prone ?.

→ Inferential stats
hypothesis testing
↓
population ✔

① Data type

(i) Quantative Data (Numerical)

(ii) Data (Categorical)

ii) Quantative Data

iii) Qualitative Data (Categorical)

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

- Zip Code → Categorical
- Age → quantative
- Income → quantative
- Marital Status (Single, Married, Divorced etc.) → Categorical
- Height → quantitative
- Letter Grades (A+, A, A-, B+, B, B-,...) → categorical
- Travel Distance to Work → quantitative
- Ratings on a Survey (Poor, OK, Great) → categorical
- Temperature → quantitative
- Average Speed → quantitative

Zip code → 1

Categorical 2. group

30450 4
30 2021
30 2023
30 30 21
50 50 20
30 2021
92603 45
34374 8

numerical data Quantitive 2. ⇓ mathematical stands summation zipcode 2. mean of zipcode 2. min & max 2.

Age group categorical

child — 0 - 18
young — 18 - 30
middle — 30 - 50
old — 50 - 70
wisdom — 70 above

Age → 10, 20, 30, 25, 40, 47

avg age 2.
min age 2.
max age 2.
} Quantitive

Ordinal vs nominal

A, A+, A- } order
B, B+, B- → ordinal

we can not order them

Letter Grades (A,B+,B,B-)
Types of Fruit(Apple,Banana,etc.) → we can not order them
Ratings on a Survey(Poor,Ok,Great) → Great, Ok, Poor
Types of Dog Breeds(German Shepherd,Collie,etc.)
Genres of Moivies(Horror,Comedy,etc.)
Gender
Nationality
Education(HS,Associates,Bachelors,Masters,PhD,etc.) → School, HS, BS, Masters, PhD

Continous vs Discrete

Continous
Discrete
Continous
Continous
Discrete
Continous
Discrete

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

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

140

Quantitive Data Analysis

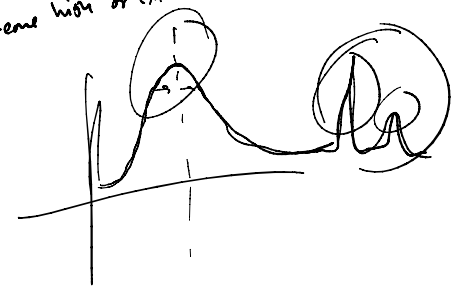
1. Measure of Center → Single Value that can describe properly of your entire data set or center point of data where or majority points in data.

Central tendency

2. measure of spread → Represents variablity in data generalize your central tendency

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

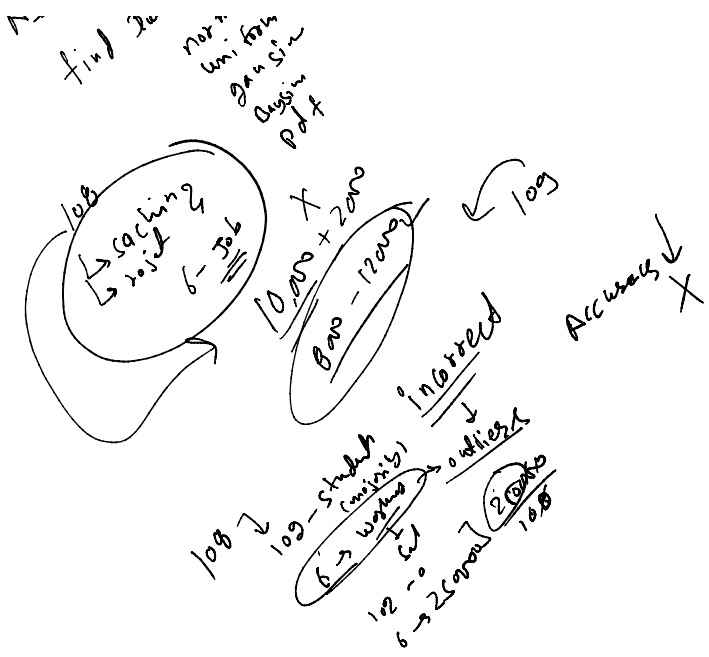
4. Outlier → extreme high or extreme low values



Assignment

find Data Distributions?

normal
uniform
gaussian



① Numpy — Vectors & Matrices

↓

Vector & matrices (Array)

Data Can be stored in →

① Scalers
② Vectors
③ matrix
④ Tensors

① Scalar is a single discrete observation
only magnitude

5
10
35.5

Each discrete observation is a scalar type

② row vector or Column Vectors

1D Array

$\vec{v}$

Vector → collection of scalars

$$v_1 = \begin{bmatrix} 10 \\ 20 \\ 30 \end{bmatrix}$$ column vector

...tor

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

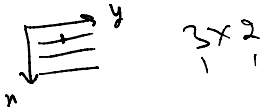
Shape → row x coll ⇒ (3, 2)

Size = row * cols ⇒ 6

row → records, tuple

cols → features, attributes

Dimension of matrix - 2D

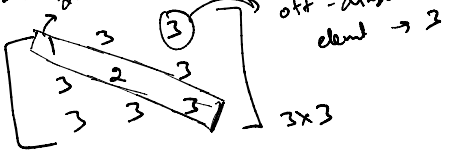


3 x 2

Square matrix

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$ 3x3

Diagonal diagonal elements [1 2 3]



d = [1 2 3]

d = 3

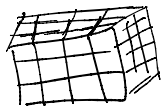
$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$ eye, Identity matrix

3x3

$$\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}_{3\times2}$$ zero matrix

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

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

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

$$\begin{bmatrix} \cdot & \cdot & 12 \end{bmatrix}$$

$$A + B = \begin{bmatrix} 8 & 10 & 12 \\ 14 & 16 & 18 \end{bmatrix}$$

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

A + B, commutative

$$A + B = B + A$$

Stats

numpy

Review

Data Science

⇒ Stats ⟨ Descriptive (Summary) ⇒ Data Analysis
Inferential ⇒ Data Analytics, machine learning

⇒ Descriptive Stats → OLAP & OLTP

Data integration transformation

ETL

Collection → Data Gathering, Data warehousing
web scraping, streaming, API, Databases, surveys

Organisation → Data cleaning, Data preprocessing, Data mining
Data Engineering, Data wrangling

Analysis → Summerization → Descriptive Stats

Visualization → Presentation → graphical, to build
into graphical, storytelling

→ Data type

① Quantitative (Numerical) ⟨ Continuous → age, height, salary
Discrete → no of movies, no of students in classroom

Qualitative/Categorical ⟨ Ordinal → grade, education
Nominal → gender, breakfast

① Quantitative [illegible]

② Qualitative / Categorical → Ordinal → [illegible]
→ Nominal → gender, breakfast

→ Data Analysis

→ Qualitative (Numerical)

Summarize / Describe

① Measure of Center (Central tendency or) → Best value that can represent the whole data → mean, median, mode

② measure of spreadness → tells about variability of data around center → range, variance, standard deviation, quantiles, 5-no summary, deciles, percentile, IQR

③ shape of data → relation of data with standard data distributions (normal, Binomial, gaussian ...) → Visualization of data, skewness, co-relation, dispersion,

④ Outliers → extreame values which does not follow characteristics of data

→ Centeral tendency

→ Let's take salary of 50 employs

X = 26, 71, 53, 96, 23, 93, 55, 87, 37, 38, 65, 75, 27, 46, 22, 23, 76, 87, 51, 32, 44, 77, 20, 97, 20, 21, 29, 63, 52, 31, 32, 35, 64, 45, 24, 84, 62, 71, 34, 60, 67, 21, 49, 32, 51, 20, 21, 77, 54, 43

So → ① mean

X = Sample Data, Random Variable

N = no of observations, size

$X = [x_1, x_2, x_3, x_4, \ldots\ldots x_N]$

$x_{10} = 38$

$x_3 = 2.$

$x_5 = 2.$

$\bar{x} = 75$

$d_{10} = 37$

$$\bar{x} = \frac{x_1 + x_2 + x_3 + x_4 \ldots\ldots x_n}{N}$$

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

Assume mean

$\bar{x} = 75$ ← Assume

d_1 d_2 d_n

$d = (\bar{x} - x_1) + (\bar{x} - x_2) \cdots (\bar{x} - x_n)$

$\therefore [x_1 + x_2 + x_3 \cdots - x_n)$

total $d_i = \sum_{i=0}^{N} d_i = \bar{x} - x_i$

deviation / scatter

if $\bar{x}$ is center than $d = 0$.

$d = (\bar{x} - x_1) + (\quad) + (\quad)$ if x_1 is ___

$0 = n\bar{x} - (x_1 + x_2 + x_3 \cdots + x_n)$

$0 = n\bar{x} - \sum_{i=1}^{N} x_i$

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

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$ ✓

$d = 0$ from $\bar{x}$

$\bar{x}_A = 75 + \sum_{i=1}^{N} d_i$

↓

$A = 1, 2, 3, 4, 5$

$\bar{x}_A = 2$ ✗

$d = 1 + 0 - 1 - 2 - 3$

$d = -5$

$\bar{x} = x_A + d \Rightarrow |2 - 5| \Rightarrow 3$

↓

$\bar{x} = \frac{1+2+3+4+5}{5}$

$\bar{x} = \frac{15}{5} \Rightarrow 3$

$\sum d = 0$

$-19 - 18 - 17 - 16 - 15 + 85$

$-85 + 85$

0

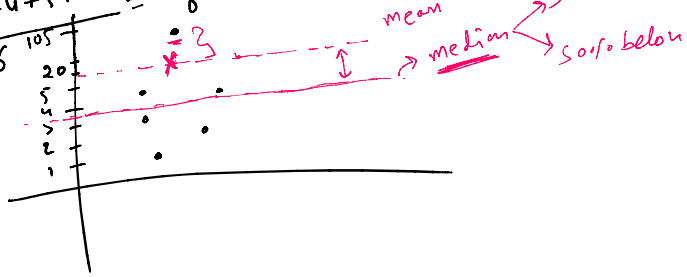
mean can be shifted by outlier

$x = 1, 2, 3, 4, 5, 105$

$\bar{x} = \frac{1+2+3+4+5+105}{6} = \frac{120}{6} \Rightarrow 20$

$\bar{x} = 20$

average of values



→ median ⇒ it always bisect data into two half

average at position

① Odd observation ⇒ median will be sorted $X[i]$, where $i = \frac{N+1}{2}$

$X = 5, 2, 4, 3, 1$, mean = 3 } uniform Normal }

$N = 5$ median = 3

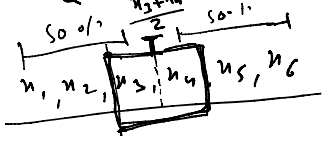
Step ① sort data (asc)

$X = 1, 2, 3, 4, 5$

median = $\frac{(N+1)^{th}}{2}$ value

$5+1 \rightarrow 3$rd value of it will be center

② Even Observation



median will be average at i^{th} and $i+1^{th}$ of sorted X data set

where $i = \frac{N}{2}$,

$\frac{n}{2}$

$X = 1, 2, 3, 4, 5, 105$

mean = 20

median = 3.5

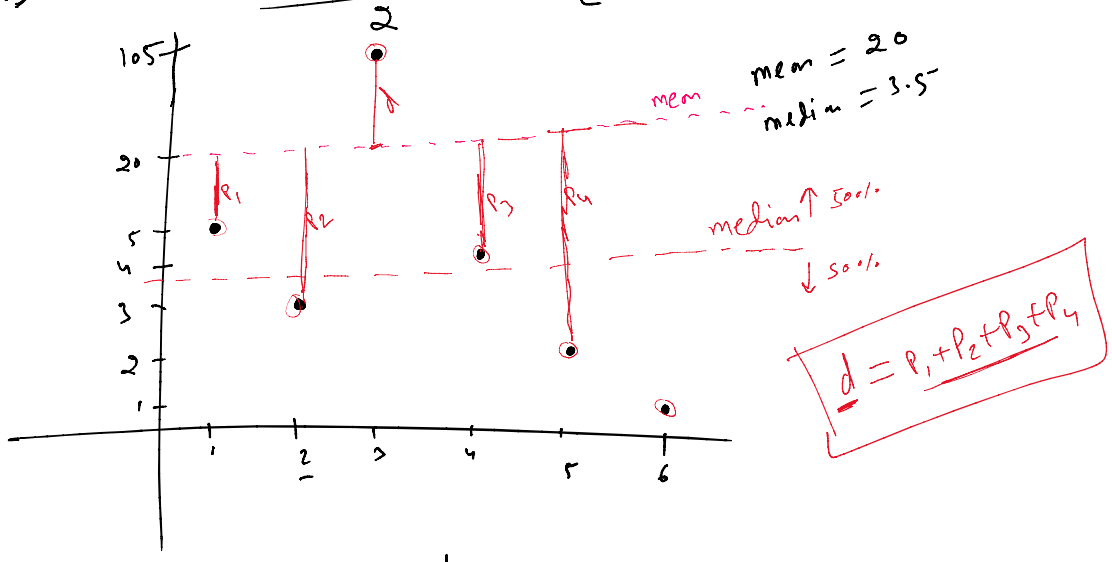
$i = \frac{6}{2} = 3$

$i = 3,$

median

$X = [5, 3, 105, 4, 2, 1]$

$\left[x_i + x_{i+1} \right] \Rightarrow \left[\frac{3+4}{2} \right] \Rightarrow 3.5$



mean = 20

median = 3.5

median ↑ 50%

↓ 50%

$d = P_1 + P_2 + P_3 + P_4$

$\Rightarrow$ Discrete Data points

mode → highest frequency

Case 1. single mode

1, 3, 1, 3, 2, 1, 3, 2, 1, 1, 5, 1, 8, 3

1	6
2	2
3	4
5	1
8	1

→ mode = 6

Case 2. multiple modes

3, 5, 2, 1, 3, 5, 2, 3, 5, 1, 4, 5, 3

X	f
1	2
2	2
3	4

mode = 3, 5

2	2
3	4
4	1
5	4

→ mode = 3, 5

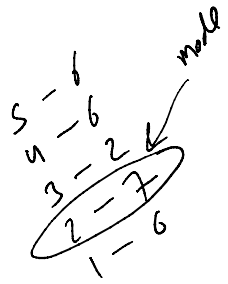
Case 3 :

No mode

2, 3, 1, 2, 4, 3, 1, 4

x	f
1	2
2	2
3	2
4	2

→ No mode



---

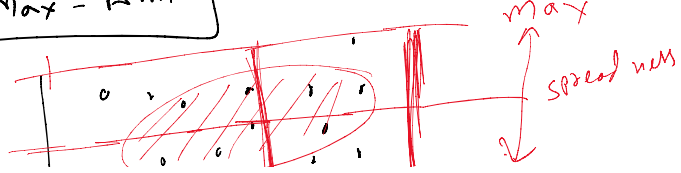
② Measure of Spread

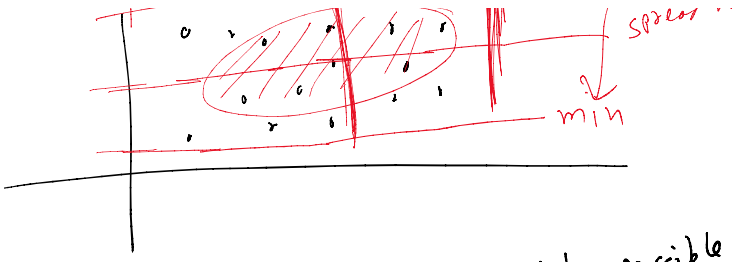
① Range
② Variance
③ Standard deviation
④ Quantiles
⑤ Inter quartile Range
⑥ 5 - point summary
⑦ Deciles
⑧ Percentiles

Distribution of Data Around Center

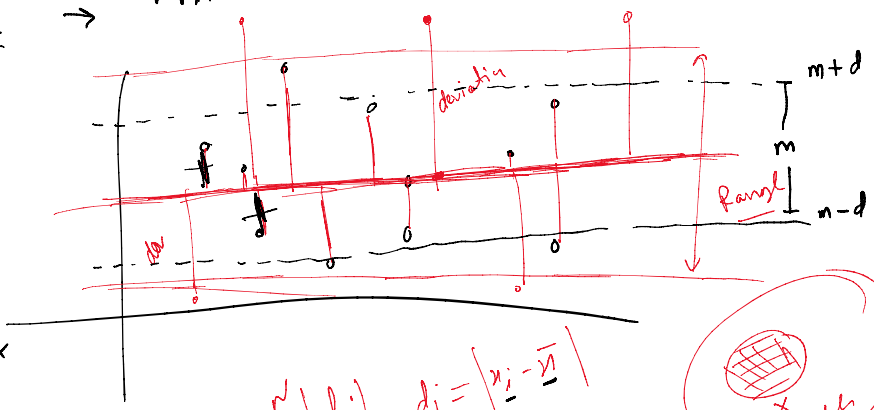
① Range → Variness of entire data set

$$R = A_{max} - A_{min}$$





② Variance → Maximum Variability possible



125 → Height

Single → mean → 5.8 feet → feet

$d = 0.2$

5.8 ± 0.2

$5.6 - 5.10$ actual

deviation = $\frac{\sum_{i=1}^{N} |d_i|}{N}$, $d_i = |x_i - \bar{x}|$

Variance → squared deviation

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N}$$

③ Standard deviation →

$$\sigma = \sqrt{\text{Var}(\sigma^2)}$$

$$= \sqrt{\frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{}}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}}$$

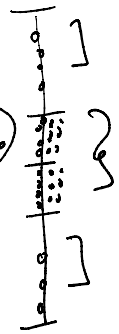
→ mean ± sd
↓ center ↓ spread distr → most common values

mean sd
→ 5.6 .4

→ mean ± sd
5.6 ± .4 →

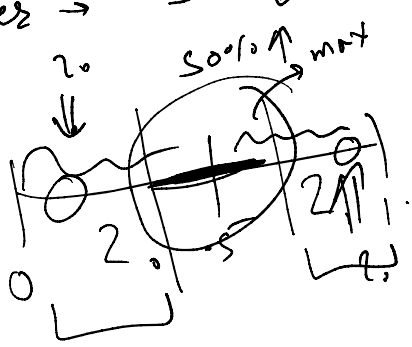
4.2
5.2
5.6
5.10
6.5

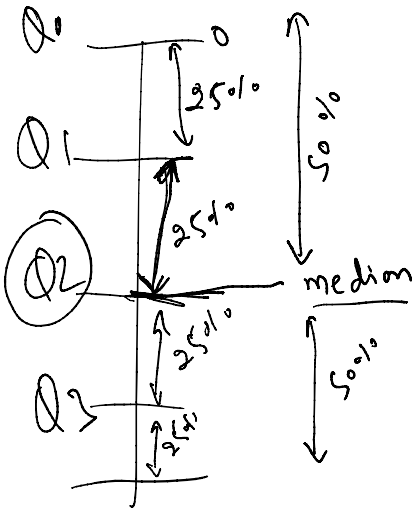
→ 30 - 70 %



④ Quantiles

median → center → 50% ↓ 50% ↑ most





data = 4, 1, 3, 8, 6, 2, 5, 7, 9, 6, 10

mean = 5.5

data = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10

Q_2 = (median) → 5th and 6th Avg

$\frac{5+6}{2} \Rightarrow 5.5$

median = 5.5

2

median = 5.5

Q_1 → median of 0 − Q_2 values

Q_3 → median of Q_2 − Q_4 vals

Q_1 = 1, 2, 3, 4, → 2.5

Q_3 = , 6, 7, 8, 9, 10 → 8

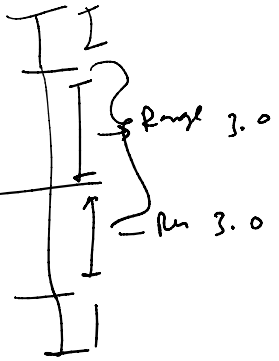
1 — 10

→ 5.5 → center , sd = 2.

→ 1st 25 % → 2.5

→ 50 % → 5.5

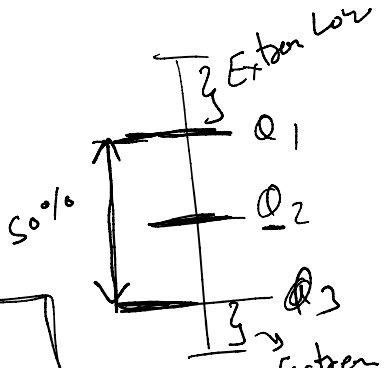
→ 75 % → 8



5.5 − 2.5 = 3.0

8 − 5.5 = 3.0

→ $Q_1 - Q_3$



(5) IQR ⇒ $Q_3 - Q_1$

(6) 5 - point summary

① min → Q_0

min, Q_1, Q_2, Q_3, max

5 point summary

Explaining of your data

Range

IQR

① min → Q_0
② Q_1
③ Q_2 → center / mean
④ Q_3
⑤ max → Q_4
⑥ sd
⑦ mean → center

[min, Q_1, Q_2, Q_3, ...]
5 point summary

median − mean
↓
significal differ
↓
outliers or skewness

→ Data Distributions

→ Shape

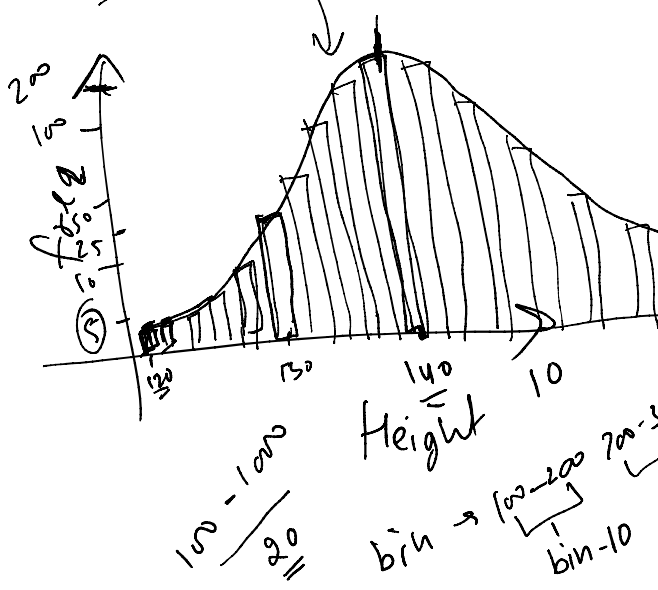
population mean = $\frac{\sum_{i=1}^{N} x_i}{N}$

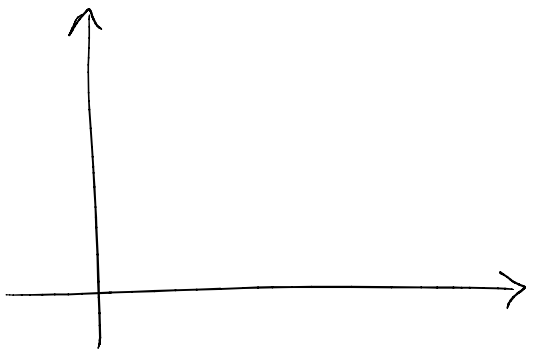
sample mean = $\frac{\sum_{i=1}^{N} x_i}{N-1}$

→ Quartiles →

(1) Histogram → Data Distribution (numerical)

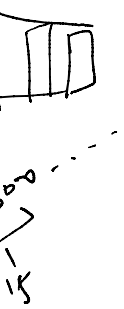
kde





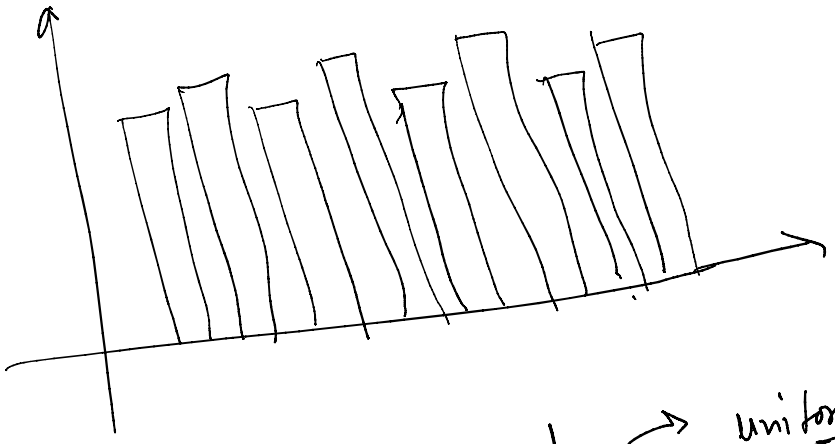
→ probability × Distribute

.. Distribution



① Uniform Distribution

where probability of occurrence of any point in data set is equal.



1 - 50

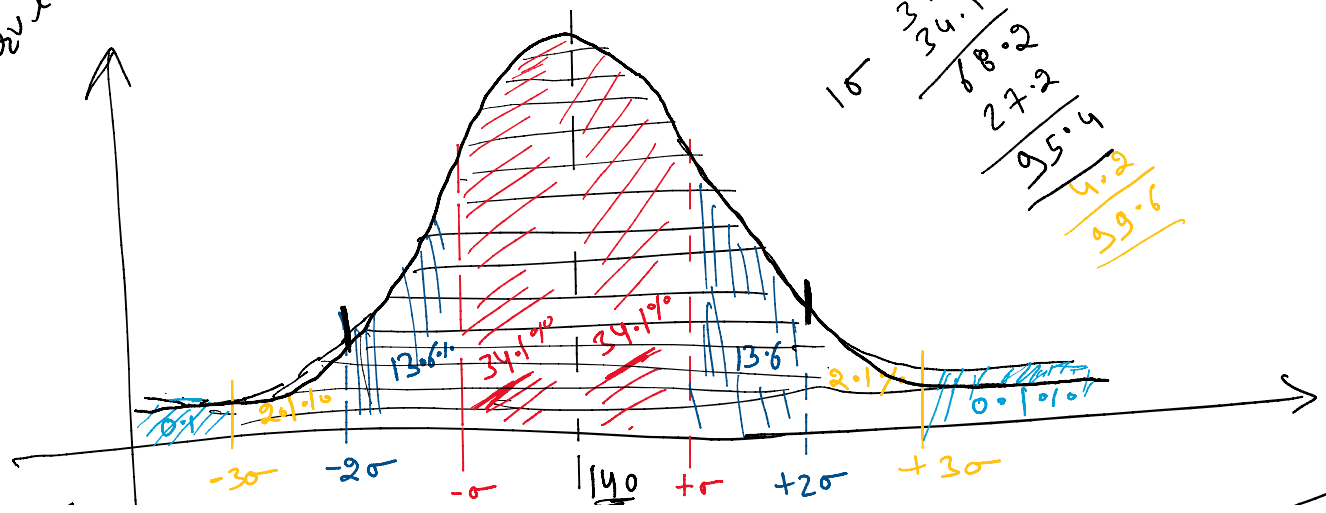
$$f(n) = \frac{1}{b-a}$$

randint → uniform

→ Normal Distribution (Gausian Distributed)

"Bell shape curve"

Symmetric



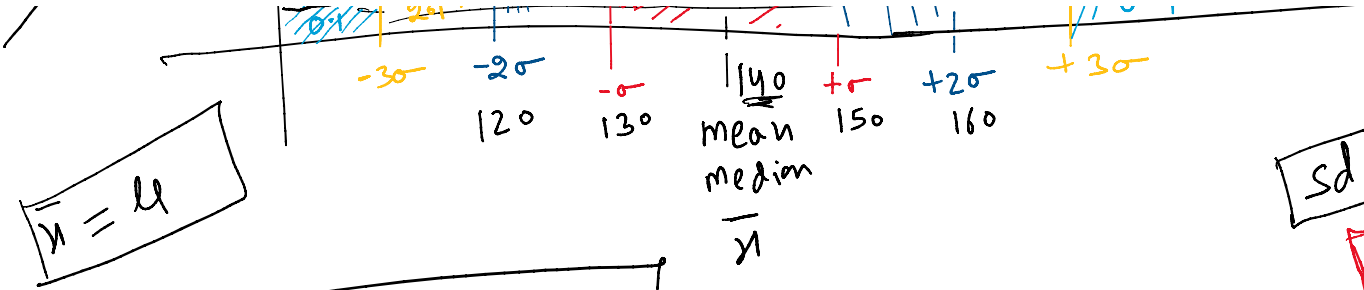
1σ 34.1 34.1 68.2

2σ 13.6 13.6 27.2

27.2 95.4

4.2 99.6

700%



$\bar{x} = \mu$

$sd = \sigma$

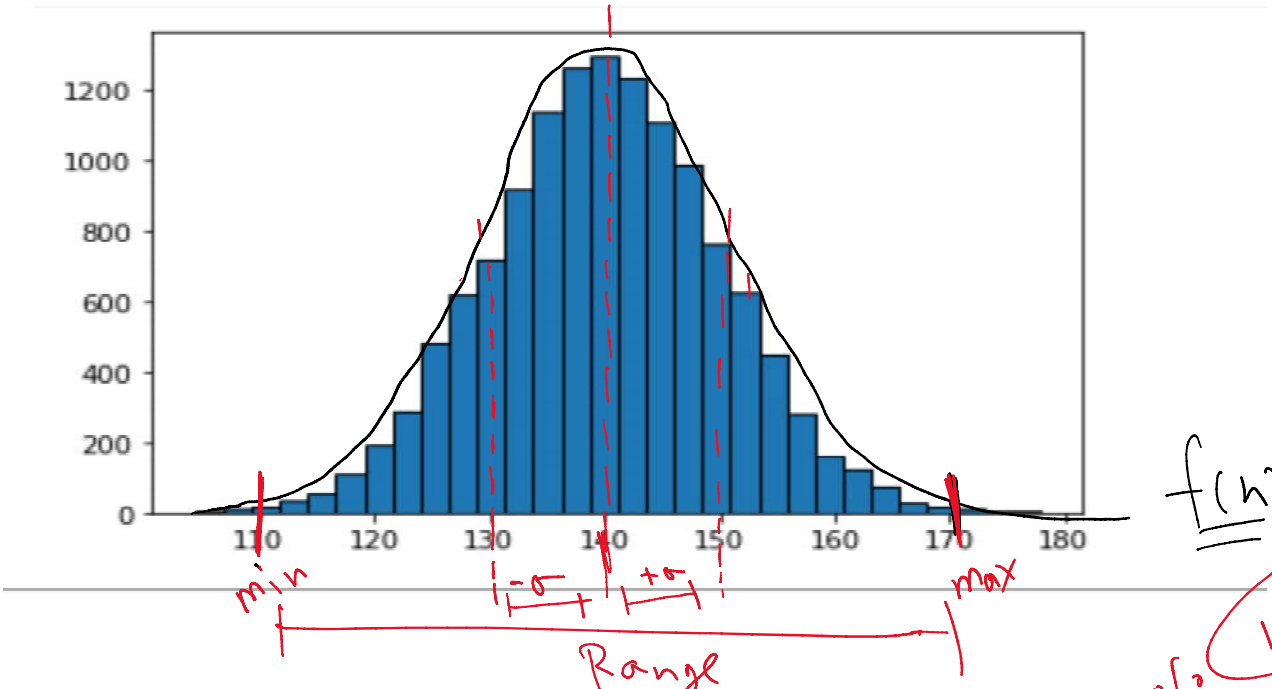
$\bar{x} \pm \sigma$

140 ± 10

$x \pm 2\sigma$

$140 \pm 20 \rightarrow 120, 160$

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



$f(x)$

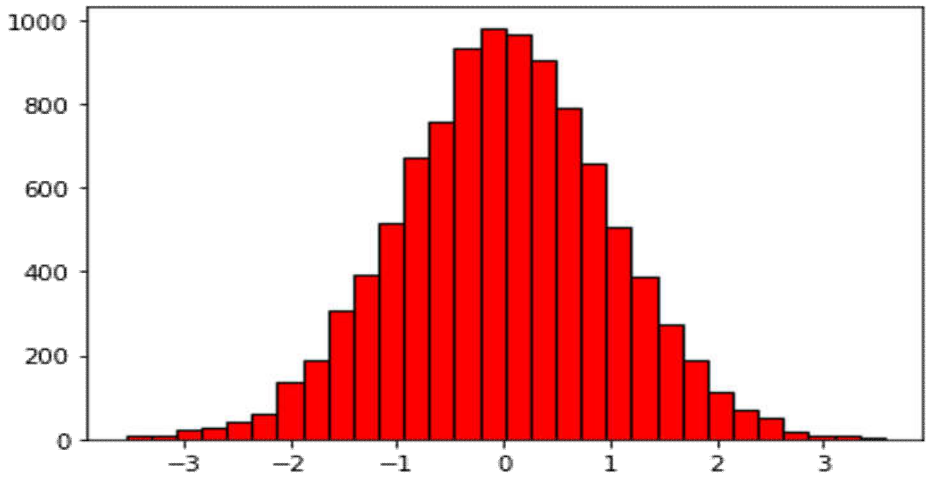
70% 130 - 150

→ Standard Normal Distribution

700%

0 - 15%

70%



mean = 0
std = 1

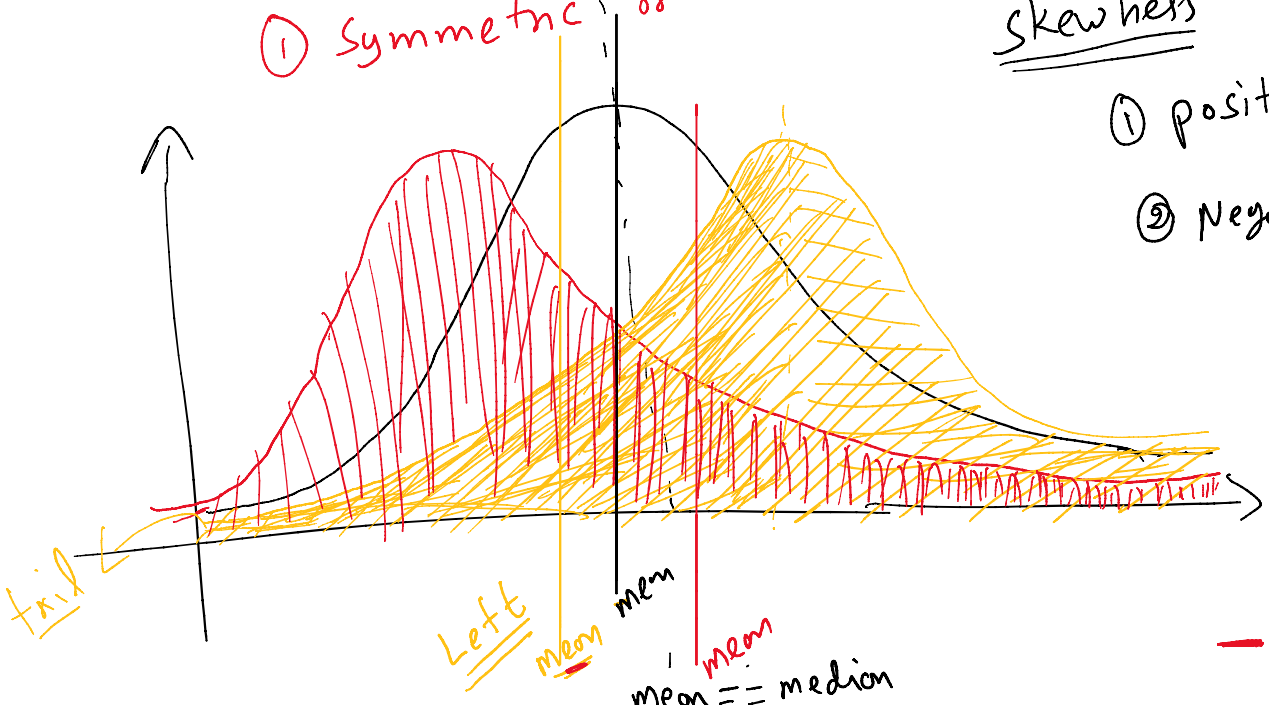
⇒ study of shape

① Symmetric or Asymmetric

Skewness

① positive skewne

② Negative skew



mean == median
(Right shift)

① positive skewness
mean > median

② Negative skewness (Left shift)
mean < median

③ Normal or Symmetric
mean == median

- ☑ pos
- ☐ neg
- ☐ pos
- ☑ neg

⇒ Quartiles →

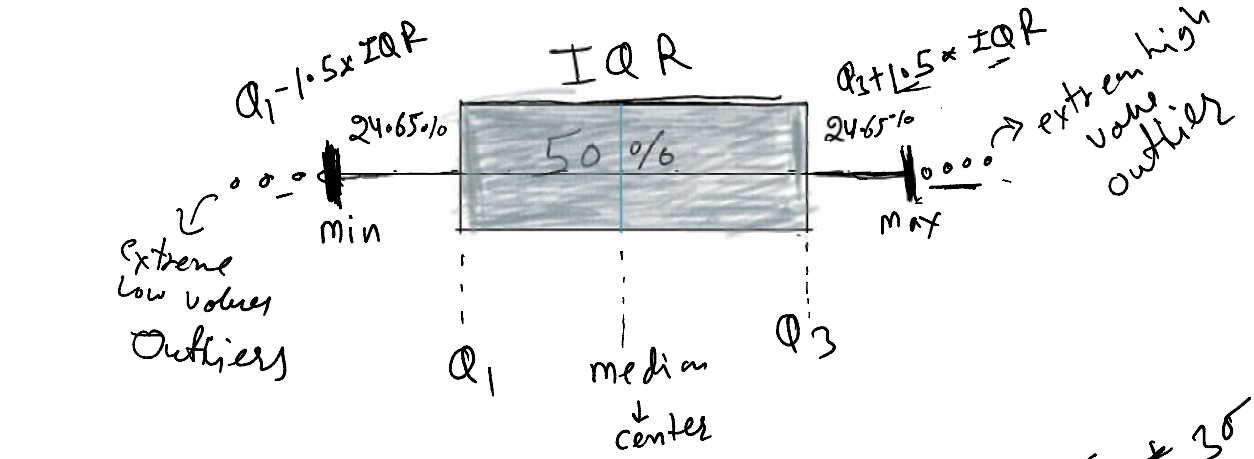
Q_1 – 25%

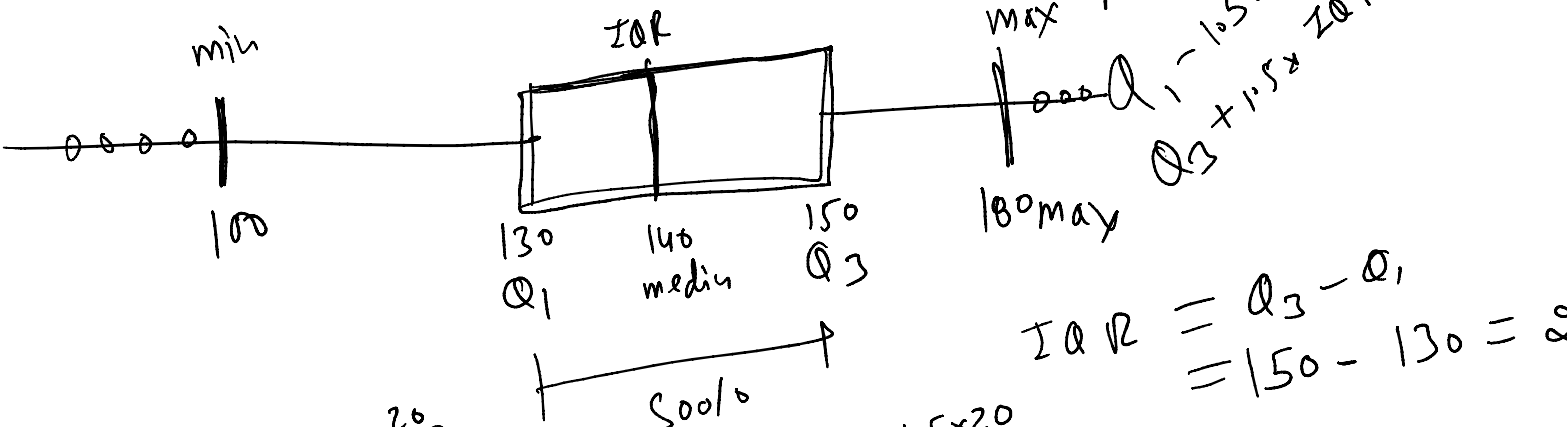
Q_2 – 50%

Q_3 – 75%

Q_4 – 100%

Box-plot





$IQR = Q_3 - Q_1$

$= 150 - 130 = 20$

$IQR = 20$

$Left = Q_1 - 1.5 * IQR$

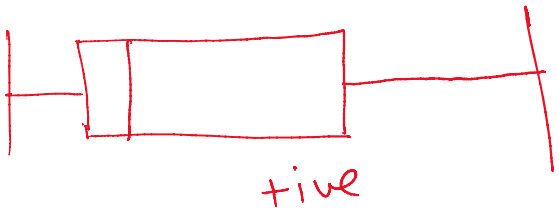
$= 130 - 30 = 100$

$Right = 150 + 1.5 \times 20$

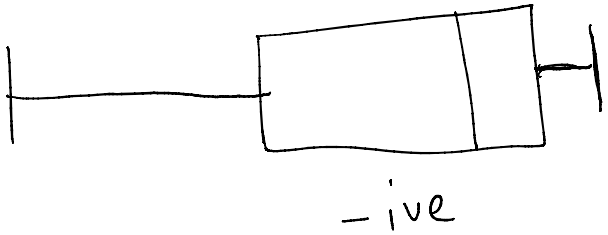
$= 180$

20

1



2



→ Summary stats

min
Q_1
Q_2
Q_3
max
std
mean

box plot
histogram

Box