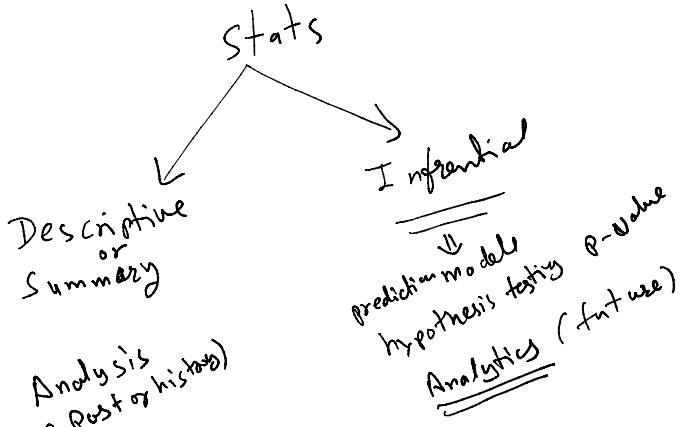
Day1_Stats

Friday, June 5, 2020 9:13 AM



⇒ Data type

① Quantative (Numerical)

② Qualitative (Categorical)

Quantitative (Numerical) → Analysis → Continous or Discrete

Center
Spread
Shape
Outliers

→ Analysis → summary

30 employ's salary

31, 86, 28, 77, 88, 34, 60, 32, 66, 52, 83, 64, 90, 37, 41, 44, 37, 69, 70, 34, 87, 45, 37, 80, 29, 83, 41, 37, 34, 45

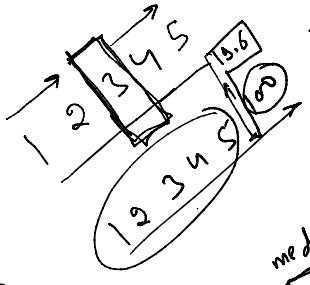
① Central tendency or measure of center → mean, median & mode

Continous mean = 54.7/2. ← 30

median = 3.5

D1 → mean=3, median=3

mean = 19.6, median = 3.5 → 2.



mean 115/6 → 19.6

19.6

median

1 4 5

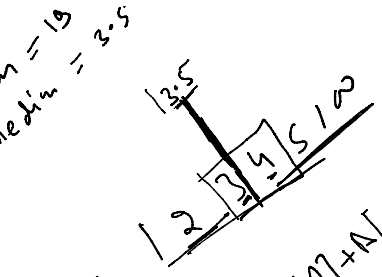
5+1/2

D1 → median

mean = 10

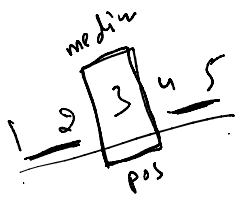
median = 3.5

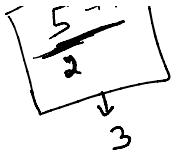
D2 →



$A = \frac{A\left[\frac{n}{2}\right] + A\left[\frac{n+1}{2}\right]}{2}$

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





→ Discrete

Q. I want to represent centre of Indian people salary?

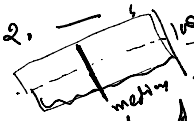
1 billion

1. — ; reason outlier
2. — ; reason outlier

Billionaires & outliers majority

mean ↗ Bias

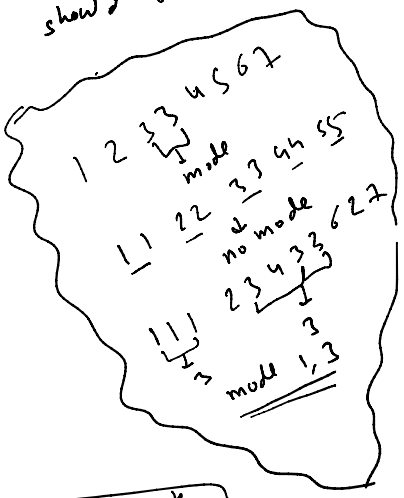
median ↙



Q. I want to summarize list of my classmate's students in a single digit what should I use

☑ mean ?

☑ median ?



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

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

median →

① sort your values

case ① ② if even observation (N)

cal $i_1 = \frac{N}{2}$

$i_2 = \frac{N}{2} + 1$, $i_1 + 1$

Then median $= \frac{A_i + A_{i+1}}{2}$

5, 10 }

$A = \{1, 2, 3, 4, 5, 100\}$

$N = 6$

$i_1 = \frac{6}{2} = 3$

$i_2 = 3+1 = 4$

$\text{median} = \frac{A_3 + A_4}{2}$

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

median = 3.5

$i_2 = \frac{n}{2}$

then $\text{median} = \frac{A_i + A_{i+1}}{2}$

A is array of data

Case (2) if N is odd, here N is no of observation

$i = \frac{N+1}{2}$

median $= A_i$ where A_i is i^{th} element in data

eg. 1, 2, 3, 4, 5

$N = 5$

$i = \frac{5+1}{2} = 3$

$A_i = A_3 \Rightarrow$ (3)

median = 3

(1) Mode → (Discrete)

single mode

Cor (1) eg. 1, 2, 3, 6, 2, 4, 3, 6, 2, 2, 4, 2, 6, 3, 1

1 → 2

2 → 5 → mode

3 → 3

4 → 2

6 → 3

multiple mode

Cor (2) 1, 2, 5, 2, 6, 3, 2, 1, 6, 4, 1, 2, 1

1 — 4 → mode

2 — 4

3 — 1

4 — 1

5 — 1

6 — 2

no mode

1, 2, 6, 4, 2, 1, 9, 5, 4, 1

No mode

Core ③ 3, 2, 5, 6, 3, 3, 6, 4, 2, 1, 9, 5, 4, 1

3 - 2
2 - 2
5 - 2
6 → 2
9 → 2
1 → 2

All freq same so **No mode**

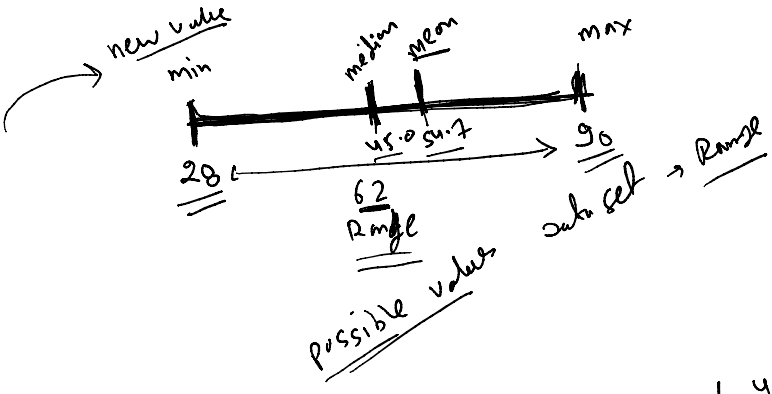
② Measure of Spread

① Range
② Variance
③ Standard deviation
④ Inter Quartile Range
⑤ 5 point Summary
⑥ deciles
⑦ percentile

① Range → maximum variability

min - max

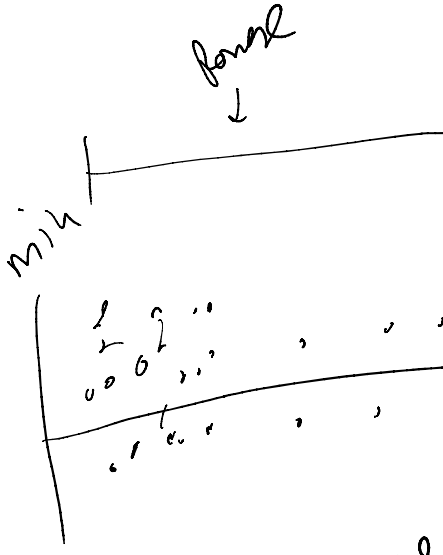
Range = max - min



50, one

possible values

subset → Range



Range

→ Summarize

max values are around 45.0
min 28 and max 90

→ Summarize max value error
min 25 and max 30

→ Variance

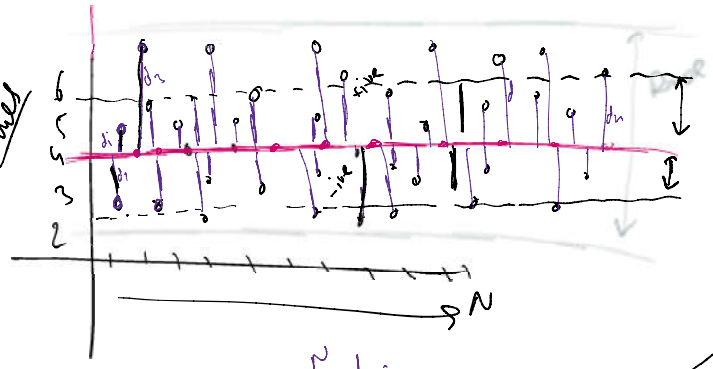
↳ it calculates maximum variance of data around center

x	$\lvert x-\bar{x}\rvert$
5	0
3	2
6	1
2	3
9	4
$\bar{x}=5$	$\frac{\sum d}{N}=2$

$x-\bar{x}$

$M \pm 2$

$$d = x_i - \bar{x}$$



total deviation $= \sum_{i=1}^{N} d_i$

$$\text{abs d} \quad \text{MAD} = \frac{\sum_{i=1}^{N} \lvert x_i - \bar{x} \rvert}{N}$$

$d^2 \rightarrow \sigma^2$

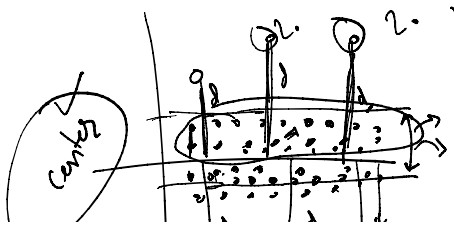
Root

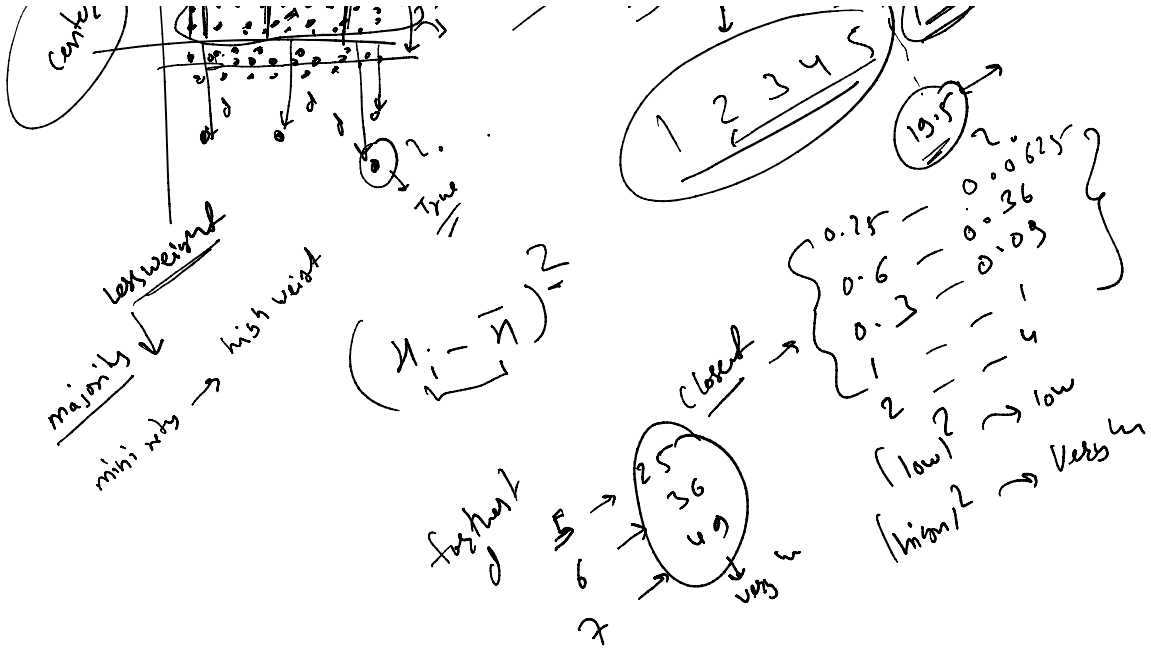
→ Variance → maximum Variability

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

$\lvert x_i - \bar{x} \rvert$ → dev / var

$\frac{\sum_{i=1}^{N} \lvert x_i - \bar{x} \rvert}{N}$ → mean → biased to majority or outliers





Variance

σ^2 → squared value

5 → 25

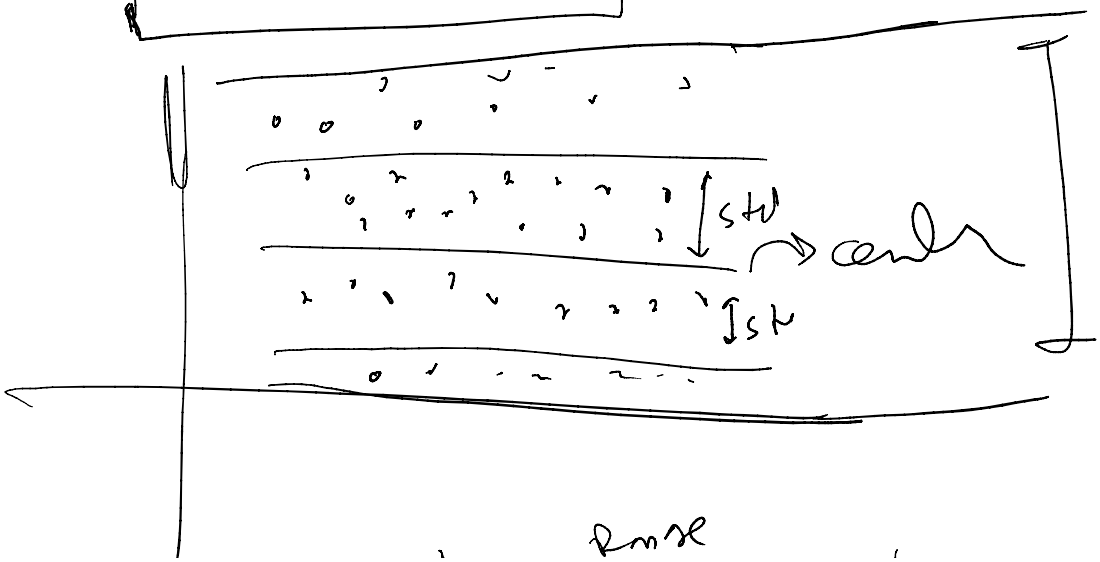
10 → 100

→ Standard Deviation

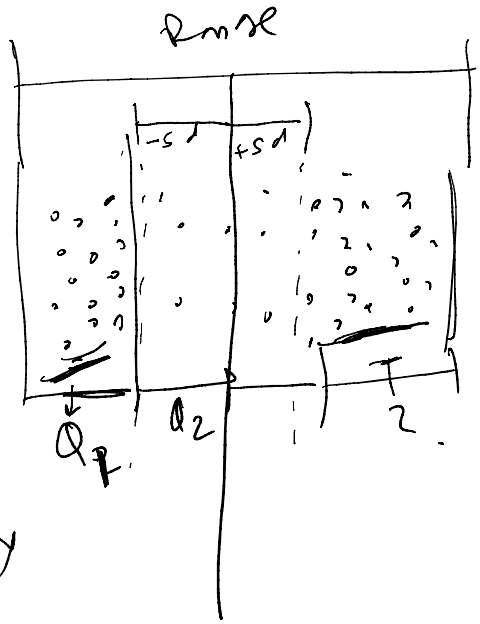
* avg. abs. dev

$$\sigma = \sqrt{\sigma^2}$$

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

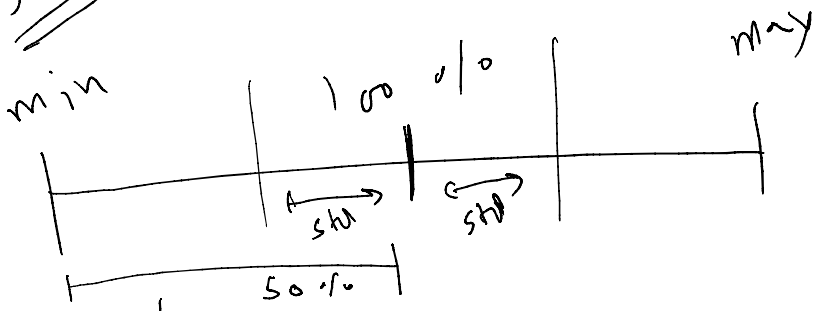


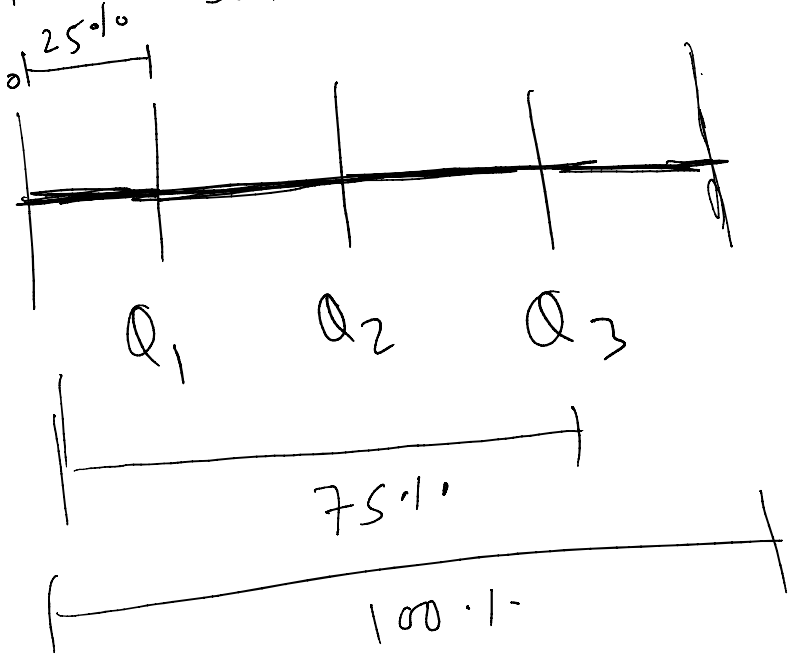
Range
std

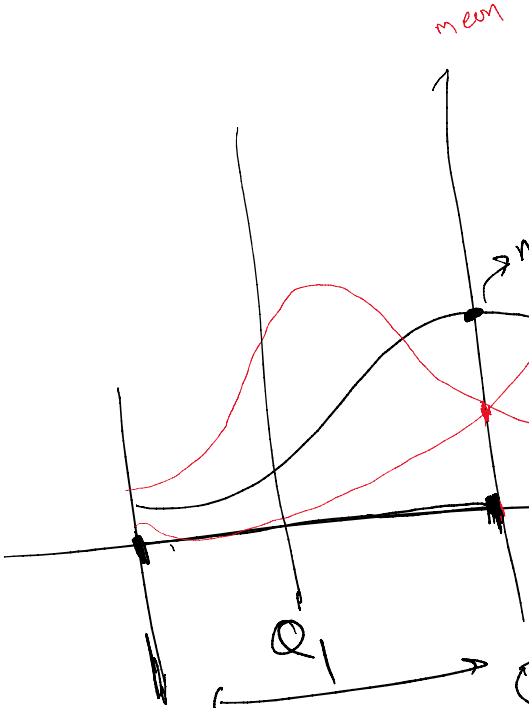


→ Quartiles

Standard

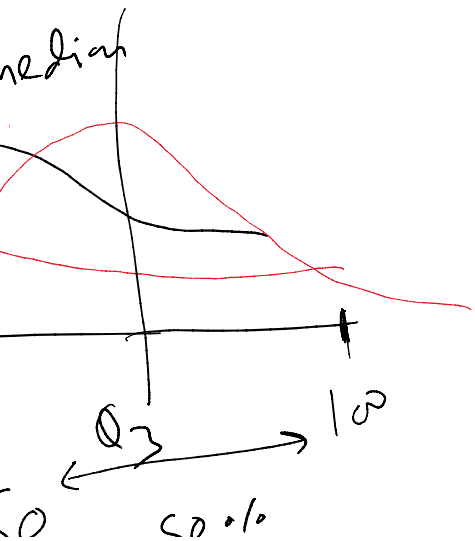






Calculation

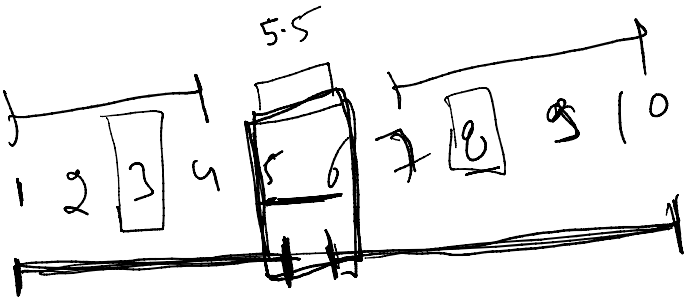
Q_2 = median


Q3
10
<10
<0.1

Q_2

Q_1 = median of min – Q_2

Q_3 = median of Q_2 – max

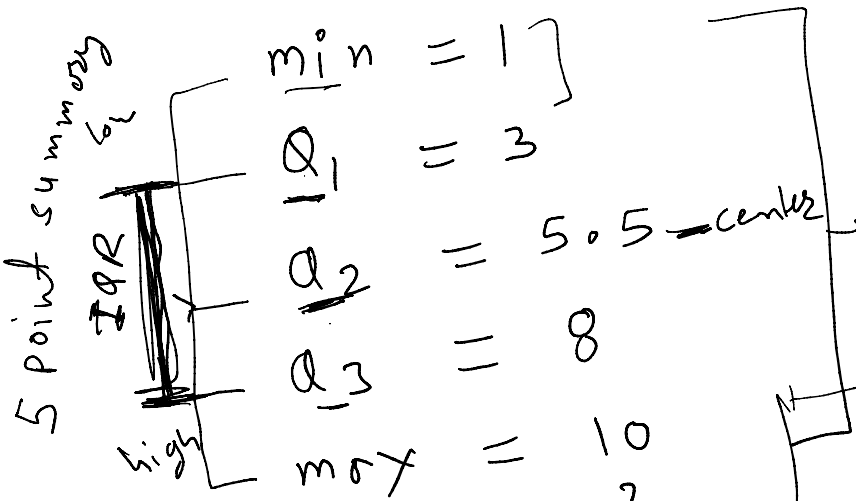


Example →

$Q_2 = 5.5$

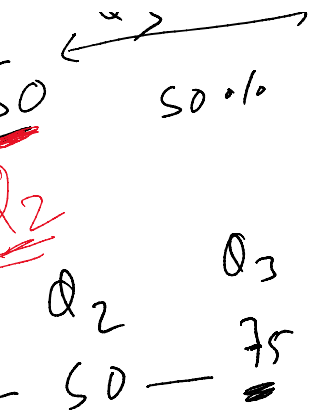
$Q_1 = 3$

$Q_3 = 8$

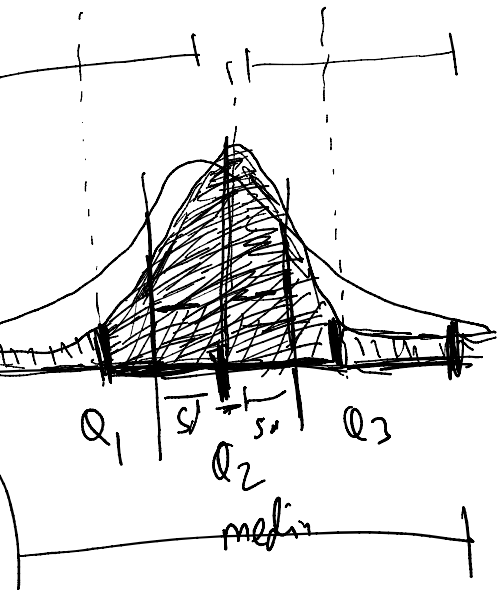


⇒ ✓ IQR → Inter Quartile Range

$$IQR = Q_3 - Q_1$$



Range



⇒ Decile

⇒ percentile

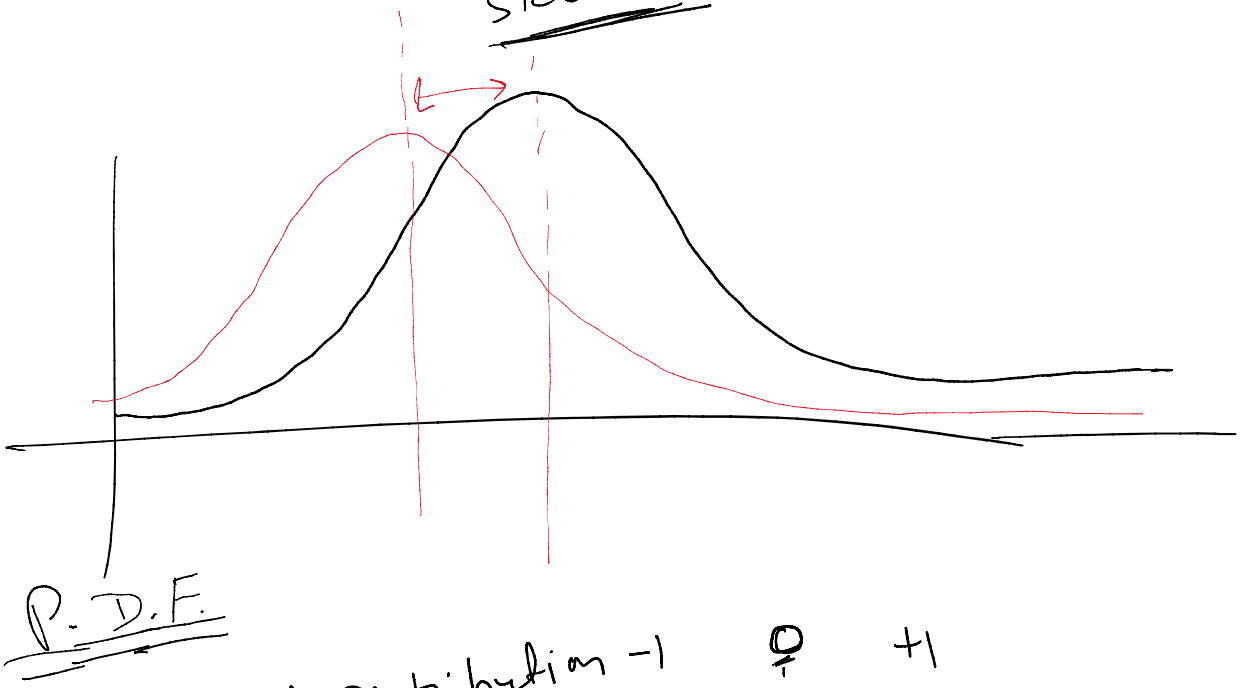
Shape of Data →

⇒ Normal Distributed

Skewness

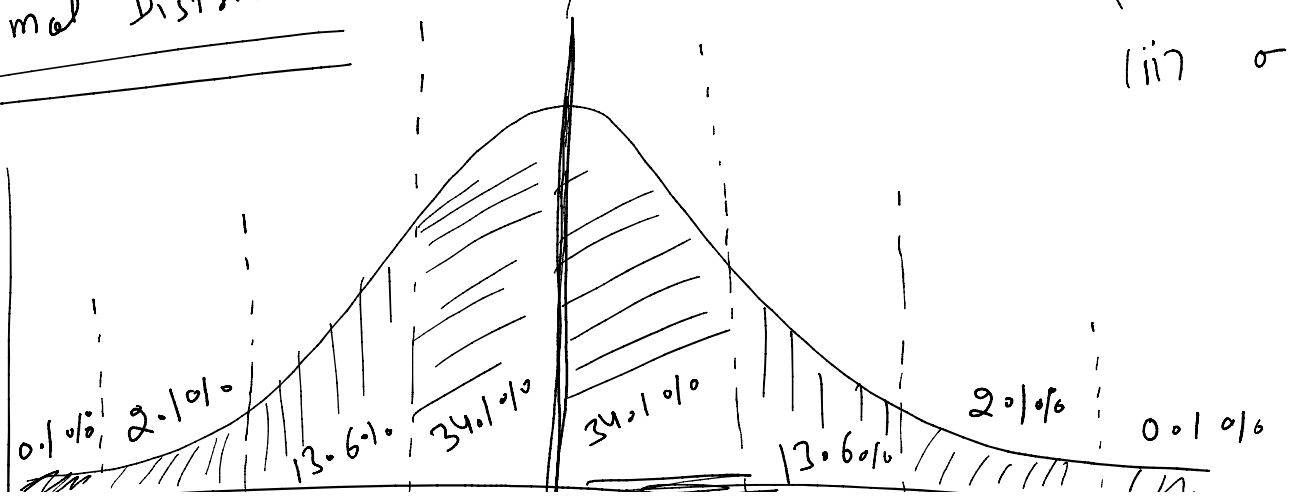
→ sym

→ A sy



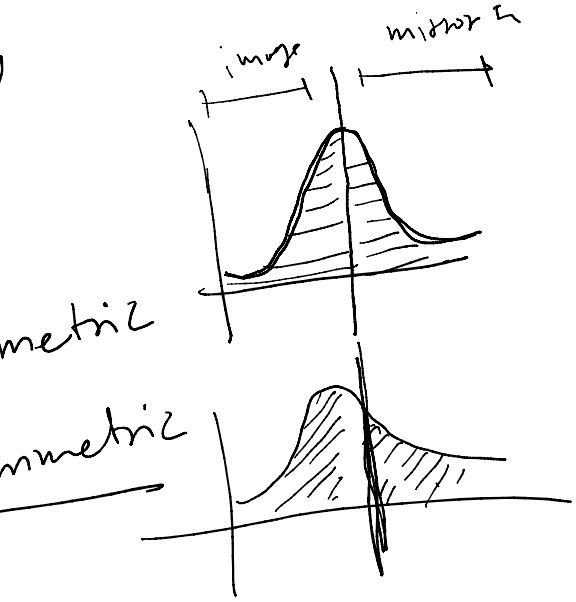
P.D.F.

① Normal Distribution



(i) mean

(iii) σ



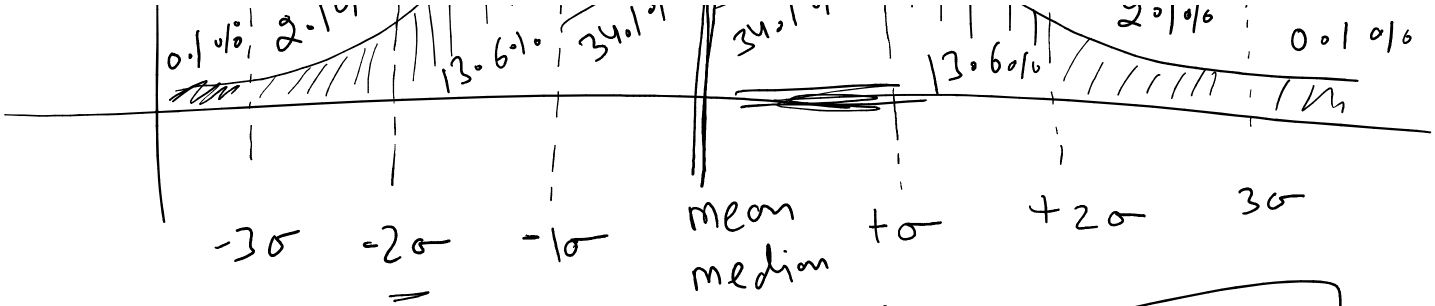
metric

mmetric

= median

std

mean = 0



70%

m ± Std → 70

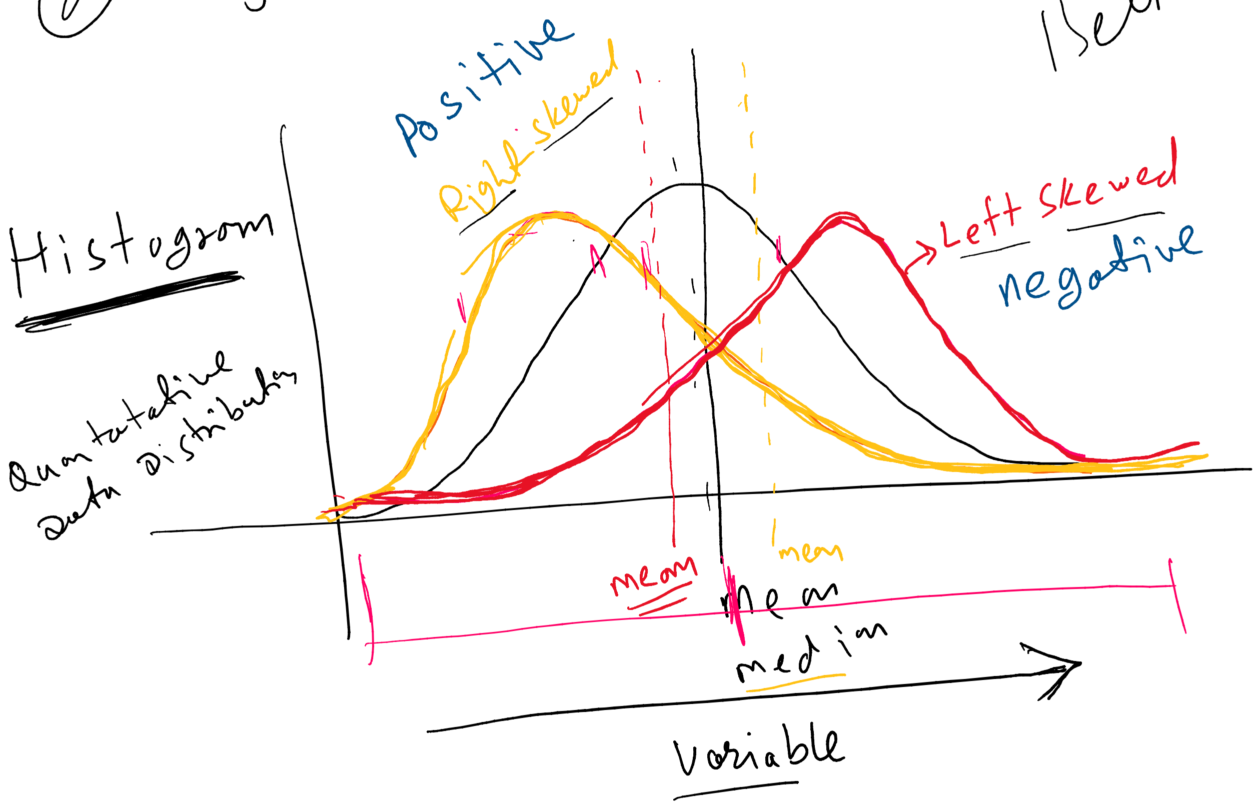
m ± 2*Std → 9

① left skewness → mean < median

② Right skewness → mean > median

Bell shaped C

Normal



mean = 0
std = 1

Standard Normal Distribution

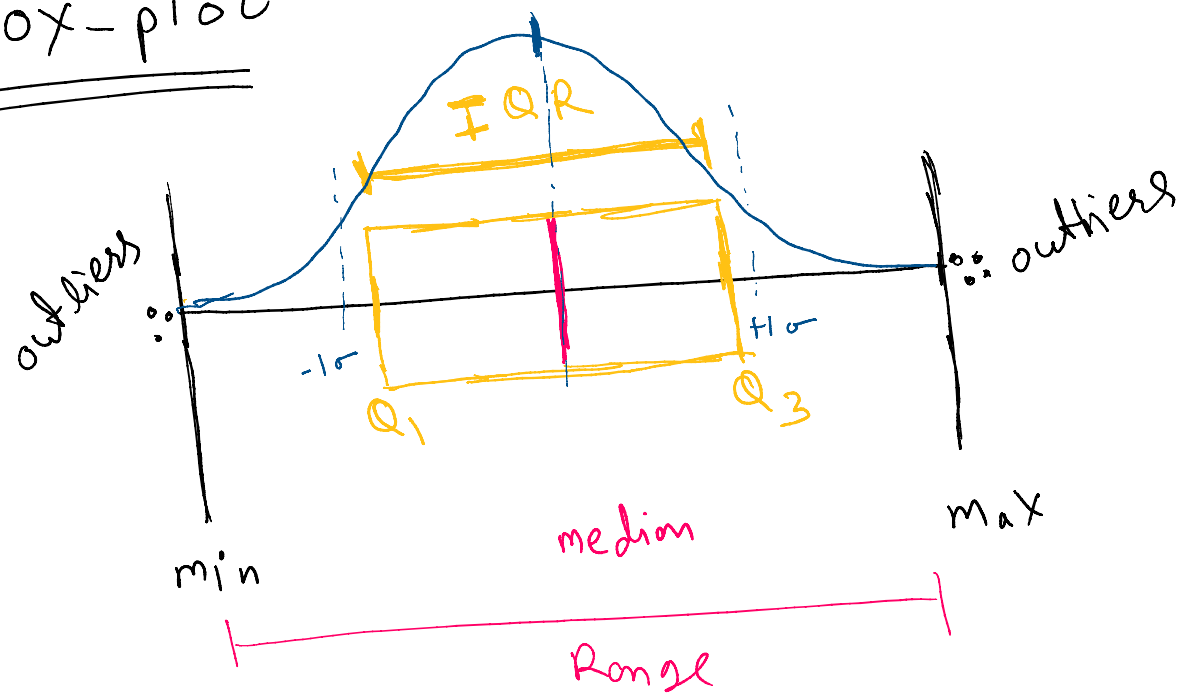
0 ± 1

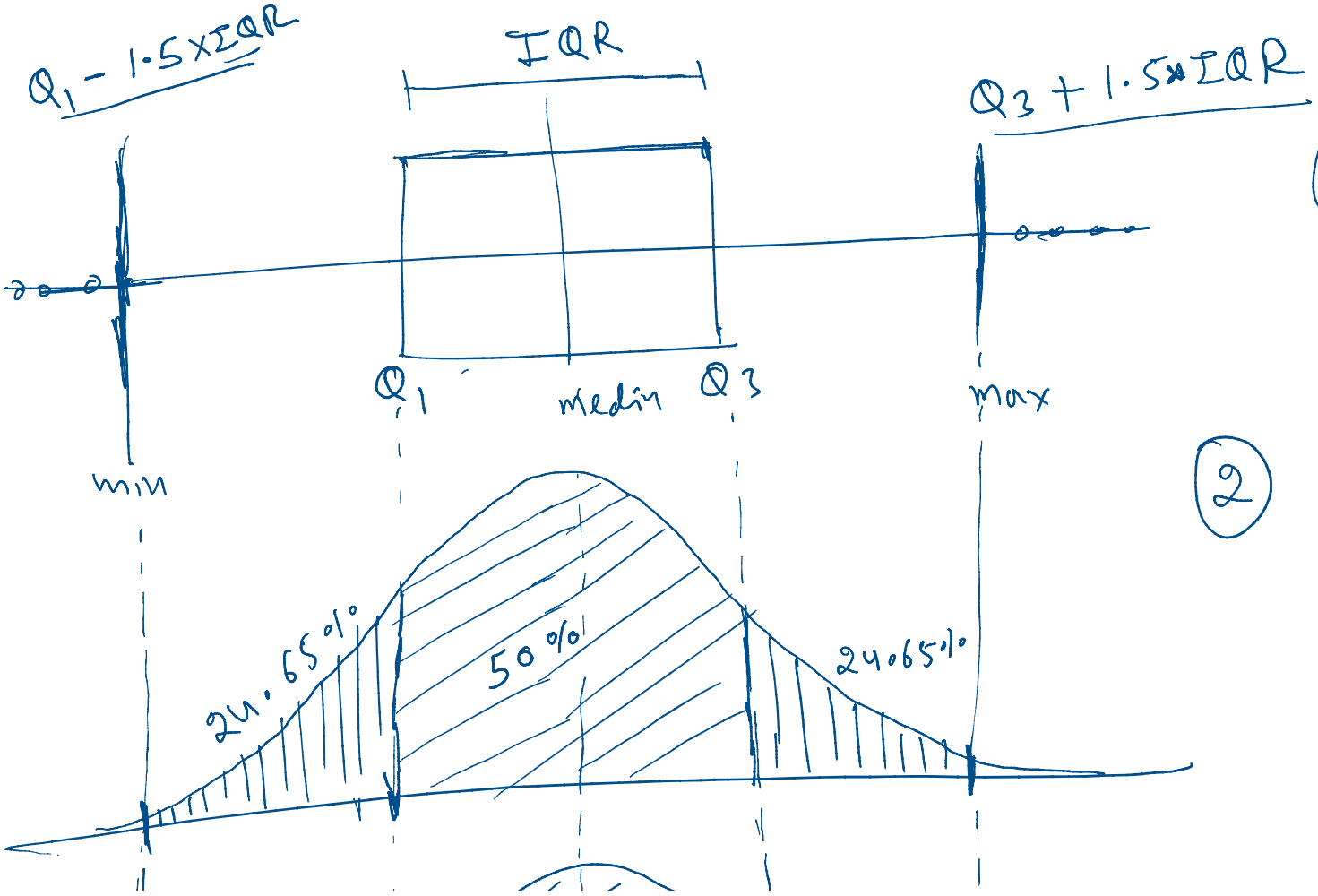
$0 \pm 2\sigma$

$0 \pm 3\sigma$

99.99

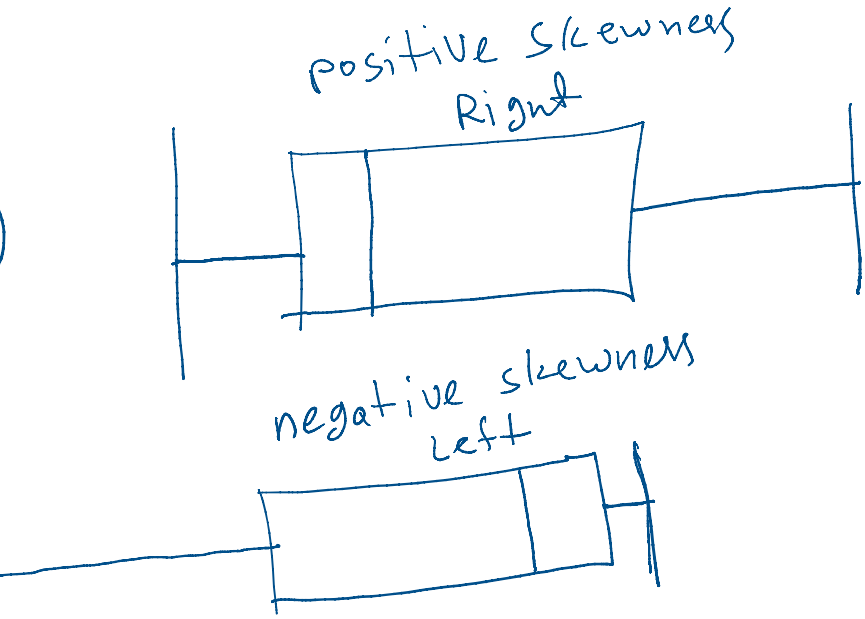
→ Box-plot →

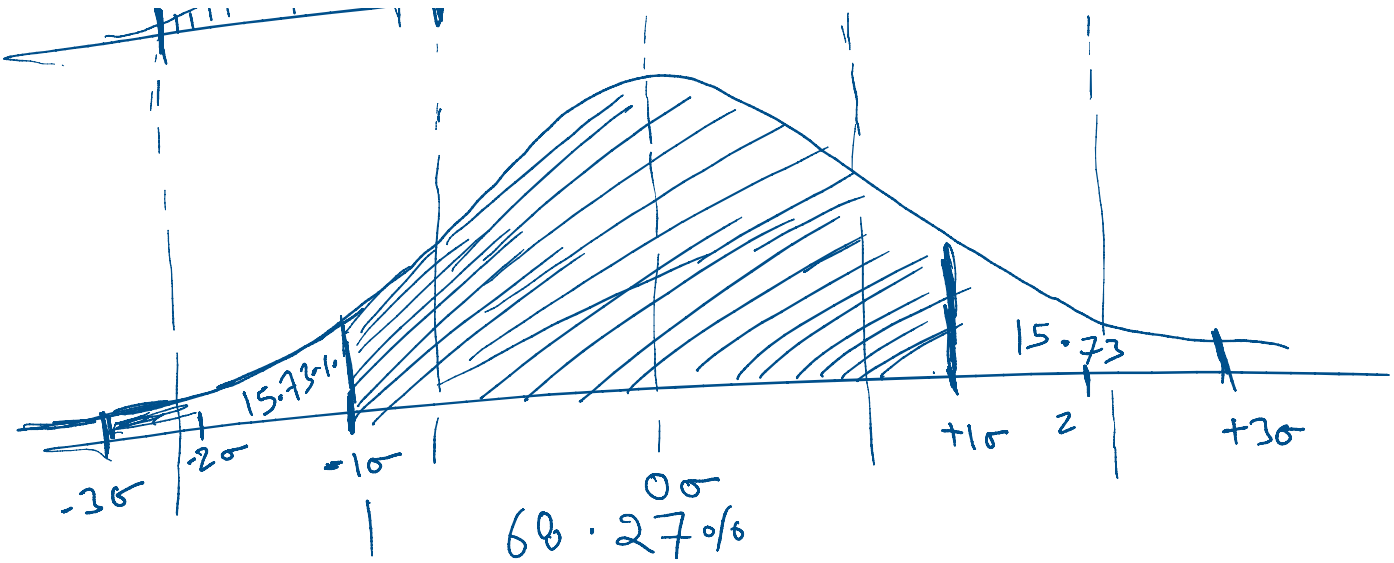

IQR
outliers
outliers
-1σ
+1σ
Q1
Q3
min
median
max
Range


Q1 - 1.5xIQR
IQR
Q3 + 1.5*IQR
Q1
median
Q3
min
max
24.65%
50%
24.65%

(1)

(2)


positive skewness
Right
negative skewness
left



⇒ Moments

$$\mu_n = \int_{-\infty}^{\infty} (x-c)^n f(x)\, dx$$

→ first moment ⇒ mean $(\bar{x})$

→ second moment ⇒ variance (σ^2)

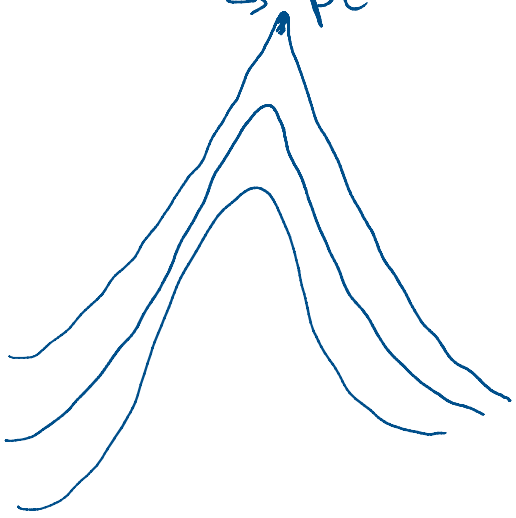
→ third moment ⇒ skewness (γ)

→ fourth moment ⇒ kurtosis

measurement

→ to get ...

↳ peak measurement



---

① Descriptive Stats → Distribution of data

② Intential stats → predictive Analysis

Population
[Sample]
stats
parameter

post → 5.4

height → [Sa

[The

① Null hypothesis (h_0)

② Alternative hypothesis (h_A)

(Sample)

(population)

mple

re is no difference At all]

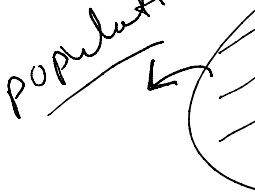
[There should be difference in data]

↳ ② Alternative hypothesis

→ hypothesis testing

→ P-value

if $p > 0.05$ we



⇒ Tests

① 1 sample proportion test

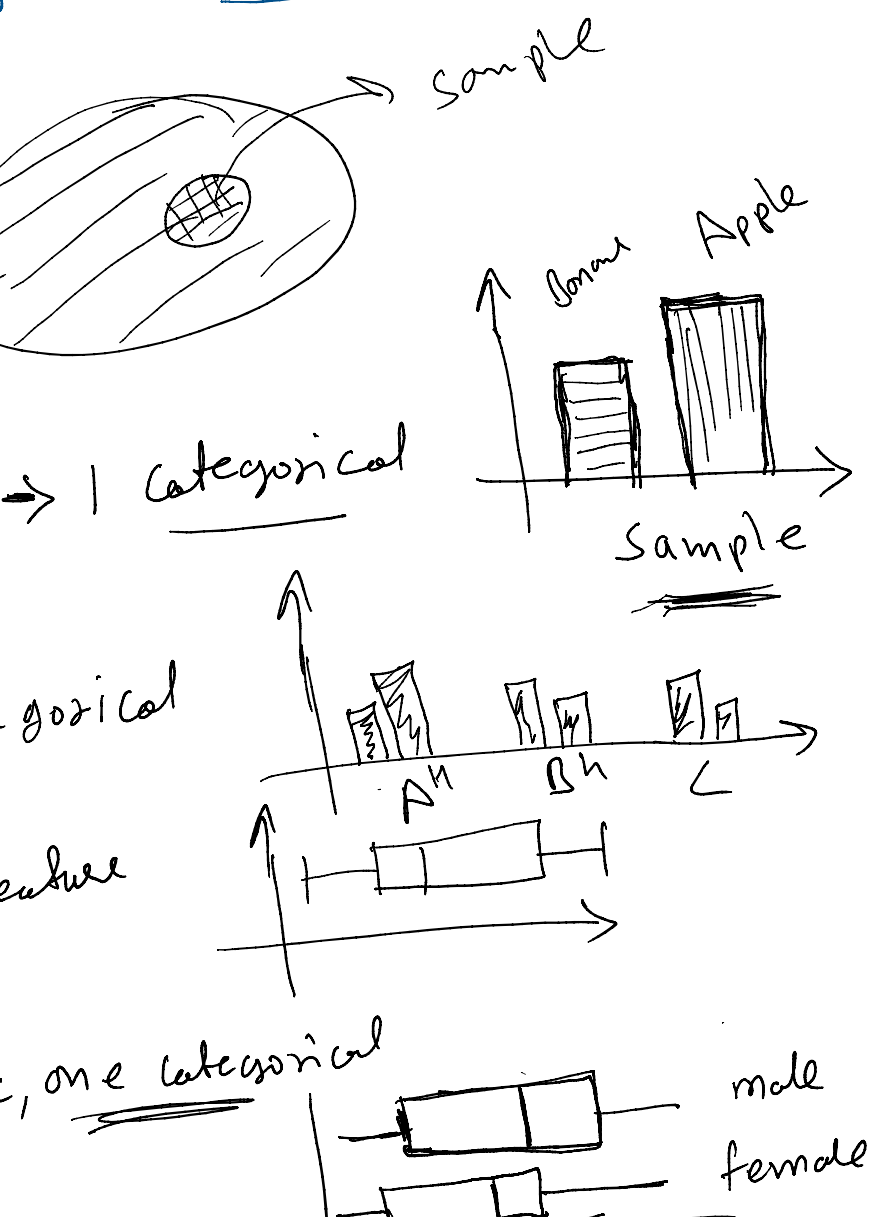
② chi-square test → 2 Cate

③ t-test → singl numerical f

test ANOVA → one numeric

$$P(h_0) + P(h_A) = 1$$

reject null hypothesis



(4) t-test, ANOVA →

(5) Correlation → 2 numerical

scatt

-1, 1

distribution [hist, Box] → numerical

corr — Scatter

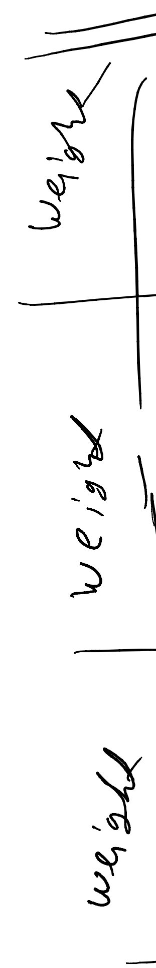
Bar, Pie → categorical

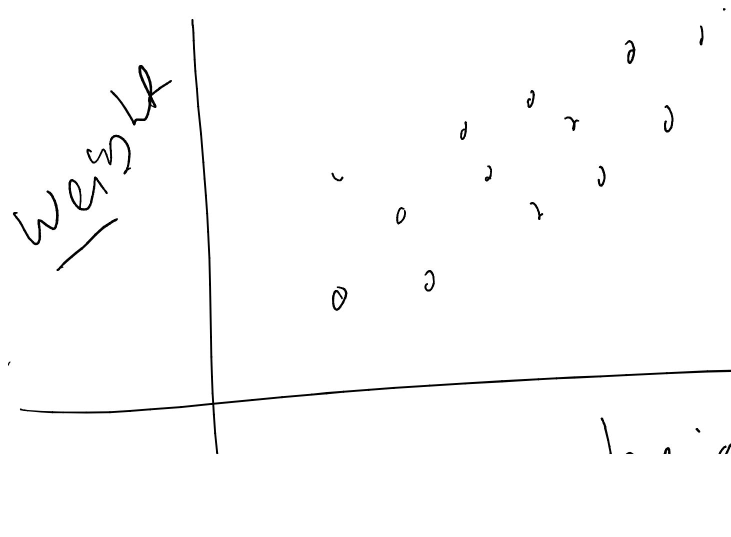
Correlation coeff → -1, 1

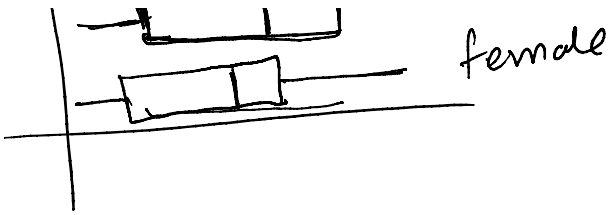
[Correlation coeff, p-value]

positive

negative







er plot

→ 0 Corr

→ +1 Corr 0-1

→ −1 Corr −1-0

Co-var

hei

→ X = height

→ y = weight

$$\sigma_x^2 = \frac{\sum_{i=1}^{N} (x_i -$$

Co-var → $\sigma_y^2 = \sum_{i=1}^{N} (y$

$$\text{Co-var } \sigma_{xy}^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})}{N}$$

Co-rr $r = \sum^{N} (x - \bar{x})$

μ

$\bar{n})^2$

N

$y_i - \bar{y})^2$

N

$(y_1 - \bar{y})$

$-\infty$ to $+\infty$

population $= N$

Sample $= N-1$

$(y - \bar{y})$

std $d > d$

std

Co-rr

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

neg pos

$$-1 \leq r \leq 1$$

$$r = \frac{Cov(x,y)}{S_x S_y}$$

$$Cov(x,y) = \sum_{i=1}^{N}$$

$$S_x = \left[\sum_{i=1}^{N} \frac{(x - \bar{x}}{N} \right.$$

$\cdot (y - \bar{y})$

$\sqrt{(y - \bar{y})^2}$

Σ

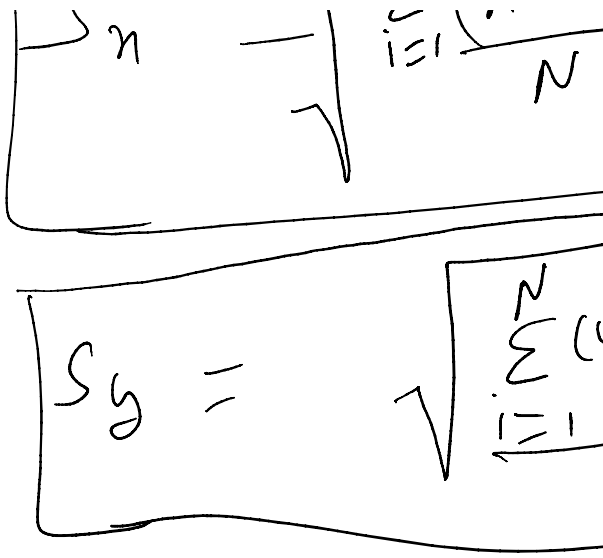
Rms(d) → Std

$\frac{d'}{\text{Std}} \rightarrow -1, 1$

$(x - \bar{x})(y - \bar{y})$

N

$\bar{x})^2$



$$s_y = \sqrt{\sum_{i=1}^{N}}$$

→ goal →

→ stats

→ matrix & vectors

$$\frac{(y - \bar{y})^2}{N}$$