STATISTICS

(1)

→ How do you decide this ATM should open here or not.

(2) Find the avg size of shark throughout world.

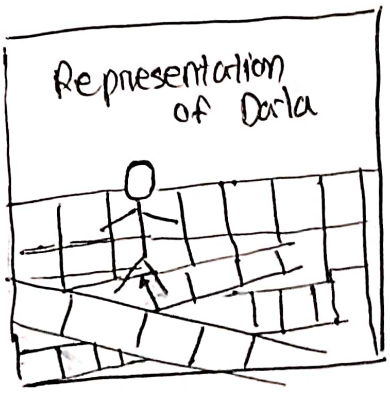
(3) Amazon Big Billion Day Sale (which month?)
{Intuit}

(4)

Life cycle of Data science project

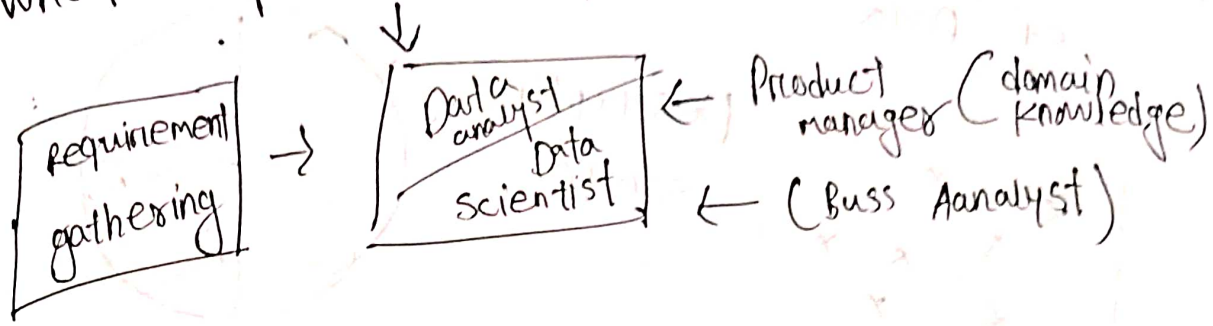
↓
product manager

"Data is more useful when we arrange data in a perfect manner."

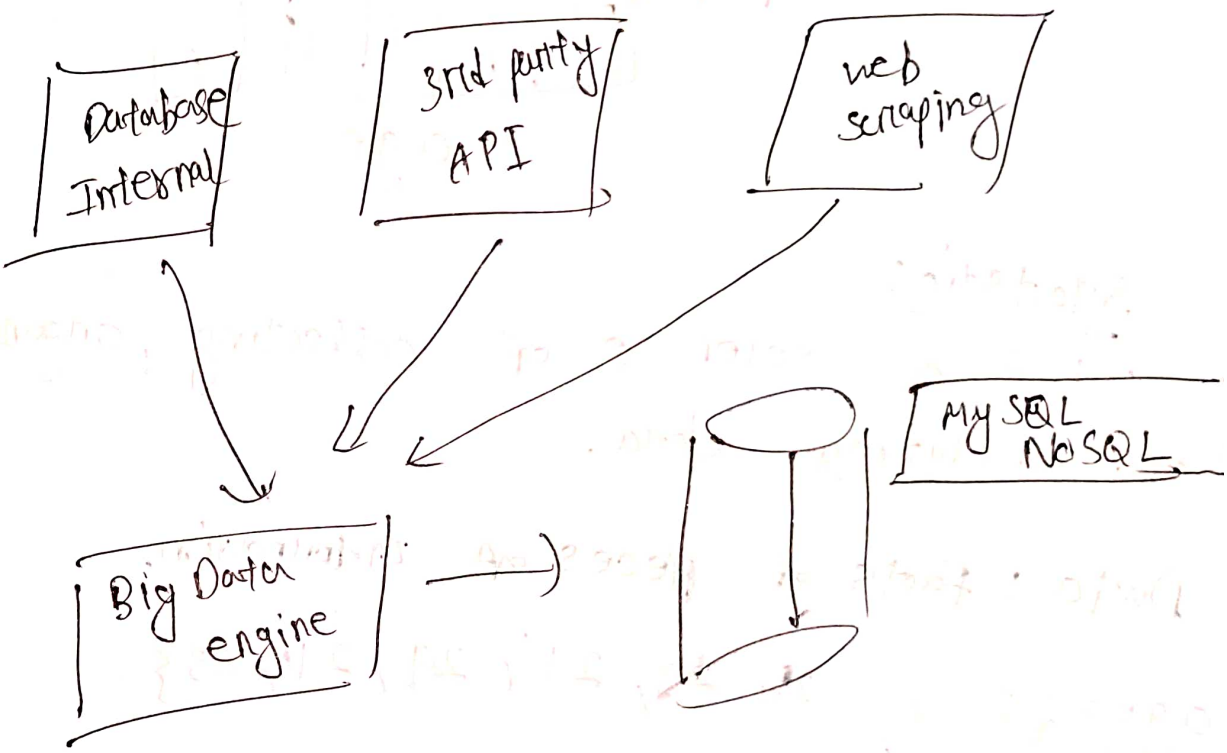


(Use of ladder)

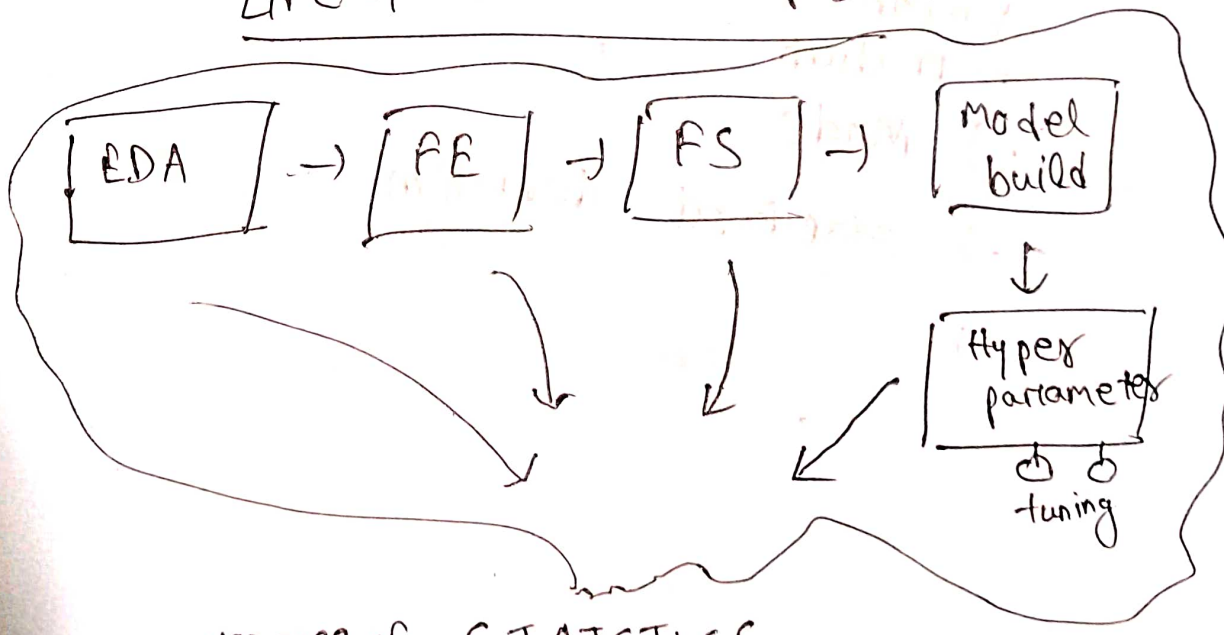
2. what companies do ?



Data ?. (collection of facts)

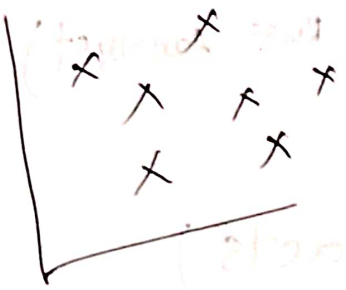


Life cycle of DS project

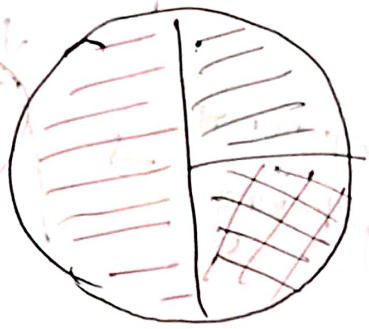


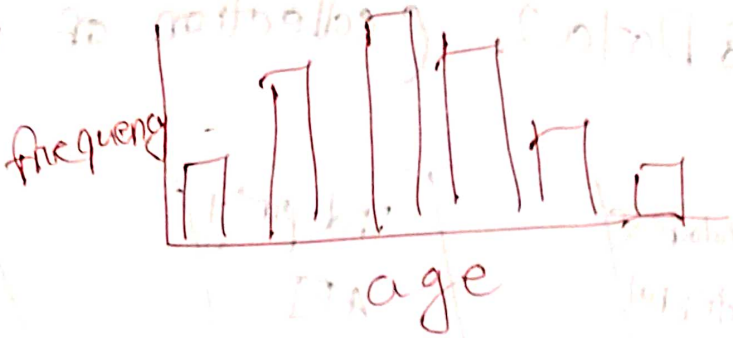
usecase of STATSTICS

Analysis of Data



Descriptive stats





Statistics

It is a science of collecting, organizing and analyzing data.

Data: facts of pieces of information.

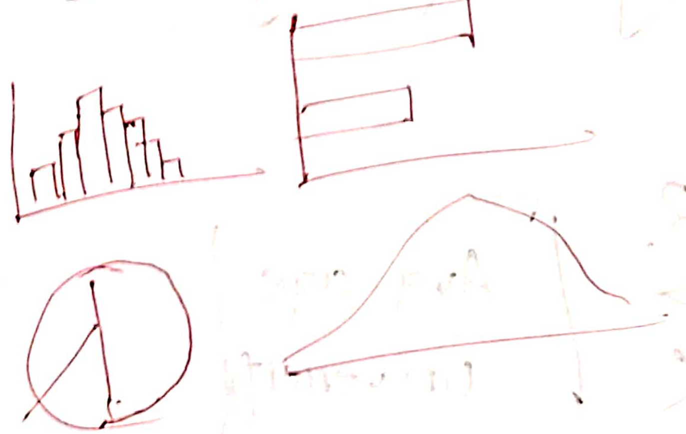
age = {21, 22, 23, 21, 24, 21, 23}

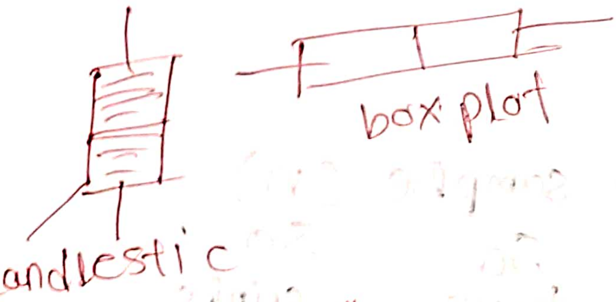
→ mean
→ median
→ Mode
→ standard deviation
→

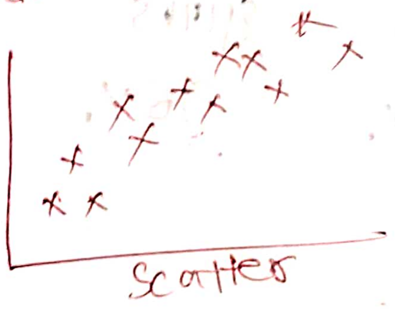
Stats

Descriptive

① Organizing, Summurizing data.



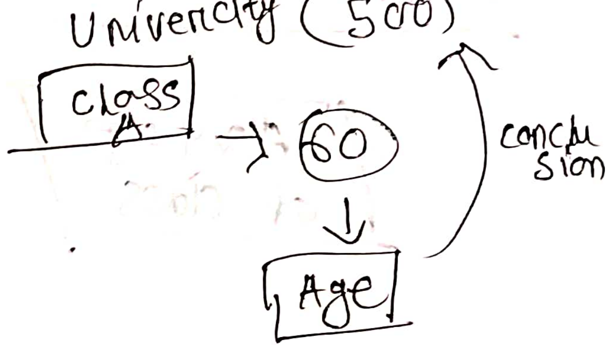


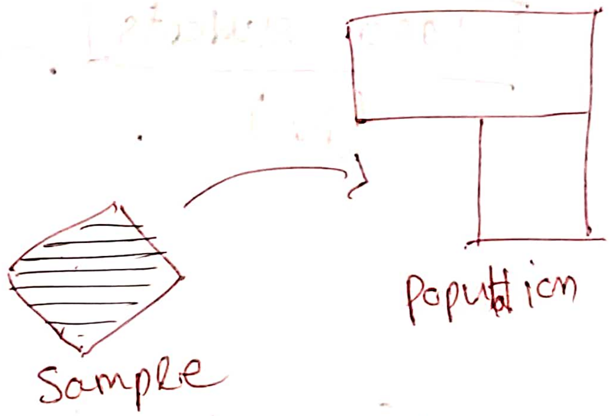


Inferential

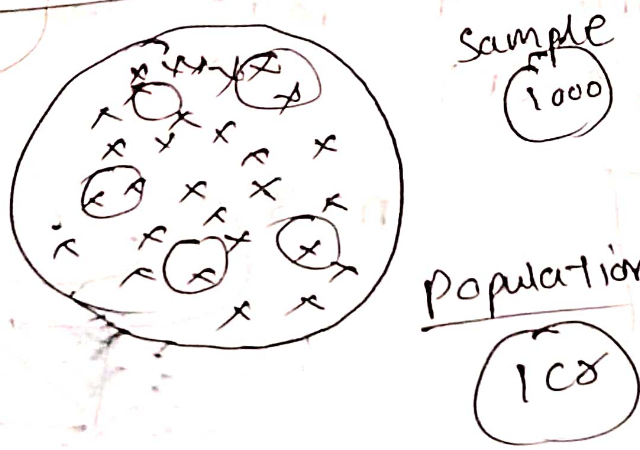
① It consist of collecting sample data and making conclusion about population.

Hypothesis testing





Sample Data vs Population Data



Eg. Age = { 21, 20, 18, 24, 28, 19, 25, 26, 23 }

weight → { ..... }

Descriptive stats

Avg → 2. age

Relationship 2. Age & weight

Inferential stats ?

Avg age of class ? Avg age University

1000 students (N) → sample (n)

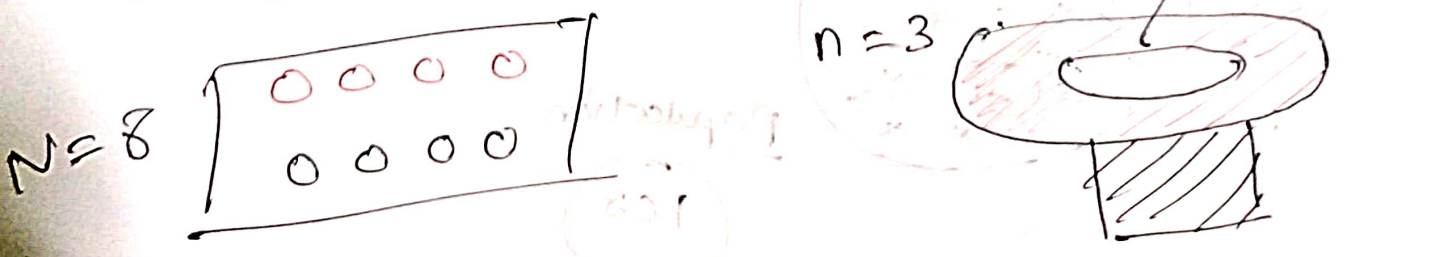
50 boys 95%.

50 girls 96%.

Sampling techniques

① Simple Random Sampling

→ Every member of the population (N) has equal chance of being selected for your sample (n)



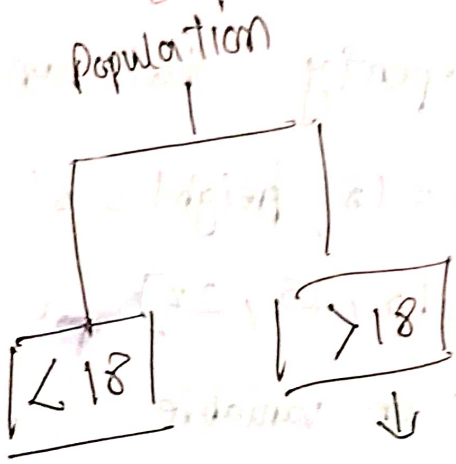
② Stratified sampling

strata → groups → layers → categories.

Gender [male, femal

Education Degree [PHD, scooling, masters

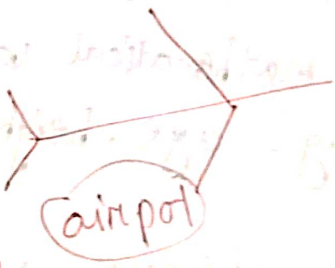
Blood groups [A, B, AB, O

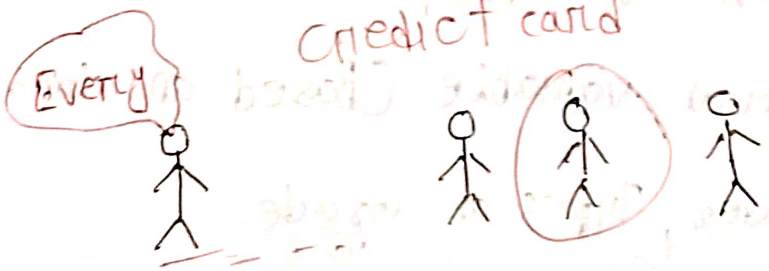


↓

Random sampling.

③ Systematic sampling :-





Selecting every nth person of a population.

④ Convenience Sampling

Only those who are intrested in the survay will only participate.

Tell me which sampling ?

→ Survay Regarding new tech (convenience sampling.

① Variable

variable is a property that can hold any value. age = 14 , age = 18 , height = 6'

ages = [13, 20, 18, 25, 23]

Two different type of variable

① Quantitative variable. →

Mathematical variable

e.g - age, height, weight, temperature.

② Qualitative variable.

categorial variable (Based on characteristics)

e.g Gender, Types of Grade.

Quantitative variable

Decrete variable

e.g whole number

Number of children, IQ

Continuous variables

weight, Height, speed

STATISTICS

Agenda

1. Histograms
2. Measure of central tendency
3. Measure of Dispersion
4. Percentile & Quartiles.
5. 5 Number summery (Box plot)

Histograms

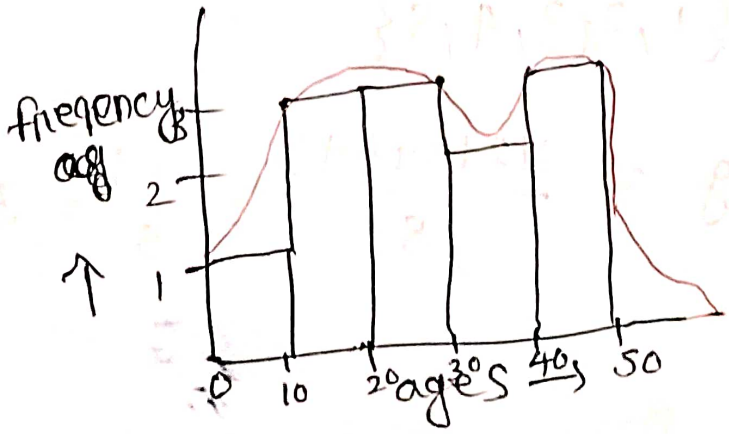
Ages = {10, 12, 14, 18, 24, 26, 30, 35, 40, 42, 43, 50}

steps

1. sort the numbers
2. Bins → No. of groups
3. Bin size → size of bins.

bins = 10

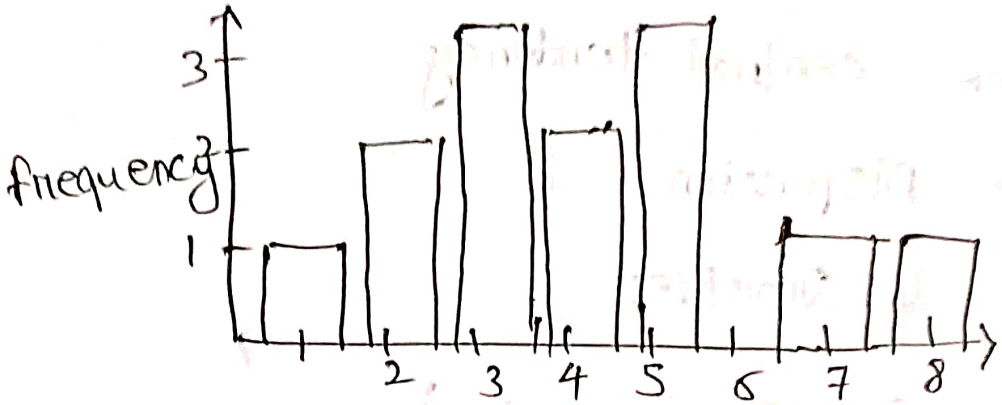
$$\text{bin size} = \frac{\text{max-min}}{\text{bins}} = \frac{50}{10} = 5$$



To smoothen this histogram we need probability density function (pdf)

② Discrete continuous

No. of Insta account = [2, 3, 5, 1, 4, 5, 3, 7, 8
3, 2, 4, 5]



To smoothen this we use probability mass function (PMS)

② Measure of central Tendency (CT)

① mean
② Median
③ Mode

→ A measure of CT is a single value attempts to describe a set of data identifying central position.

Mean.

values = {1, 2, 3, 4, 5}

$$\text{mean/avg} = \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$

Population (N)

Sample (n)

Population mean (μ)

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

Sample mean ($\bar{x}$)

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

Population age

$= \{24, 23, 21, 28, 27\}$

$$\mu = \frac{24+23+21+28+27}{6}$$

$= 17.5$

Sample age

$= \{24, 21, 1, 27\}$

$$\bar{x} = \frac{24+21+1+27}{4}$$

$$= \frac{54}{4} = 13.5$$

Practical application

Suppose you have a dataset

age	salary	family size
—	—	—
—	—	—
—	NaN	—
—	—	—
—	—	NaN
—	—	—
—	—	—
—	—	—
NaN	—	—
NaN	—	—

We can replace NAN with the central tendency.

② Median

{1, 2, 3, 4, 5} {1, 2, 3, 4, 5, 100}

$\bar{x} = 3 \longrightarrow \bar{x} = 19.16$

by adding one big value mean changed drastically.

These issues sol^n is median

Steps to find out median

① sort the numbers

② Find the central number { if no. of element odd, central element }

if, even avg of central elements.

③ Mode [Most frequent occuring elements]

e.g weather reports in a day over the time.

weather = [cloudy, sunny, sunny, cloudy, sunny, sunny]

we can say mostly day way sunny.

③ Measure of Dispersion

① Variance (σ^2)

② Standard deviation (σ)

Variance

population variance (σ^2)

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

Sample variance (s^2)

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

$\downarrow$

$\{1,2,3,4,5\}$ $\quad$ $\{1,2,3,4,5,6,80\}$

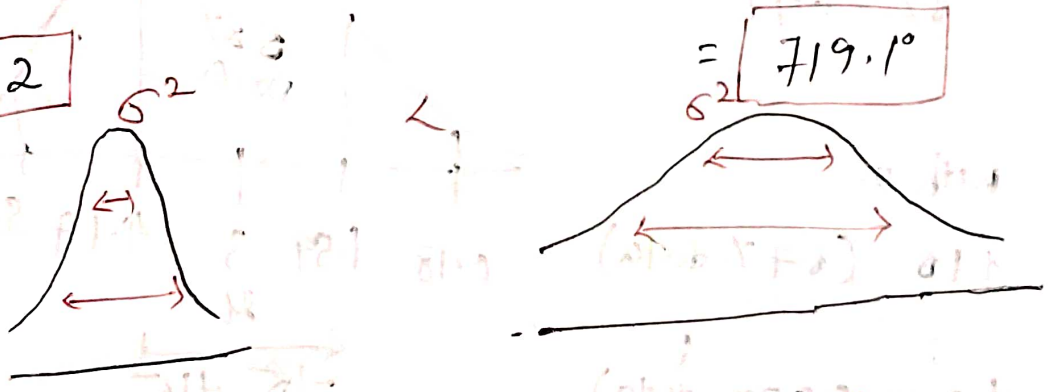
$\mu = \frac{15}{5} = 3$ $\quad$ $\mu = \frac{101}{7} = 14.4$

$$\sigma^2 = \frac{\overbrace{(1-3)^2}^{(2-3)^2} + (2-3)^2 + (4-3)^2 + (5-3)^2}{5}$$

$$= \boxed{2}$$

$$\sigma^2 = \frac{(1-14.4)^2 + (2-14.4)^2 \cdots + (80-14.4)^2}{7}$$

$$= \boxed{719.1^\circ}$$



Variance ↑↑ $\quad$ Spread ↑↑

NOTE : $(n-1)$ degree of freedom

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

sample variance is a estimator for Population variance.

When we applied to sample data, the population variance is a biased estimator for sample variance. it tends to underestimate (huge gap between sample and population variance) the amount of variability.

To solove this underestimation problem, staticians faund out by deviding (n-1).

② standard deviation ($\sqrt{\sigma^2}$)

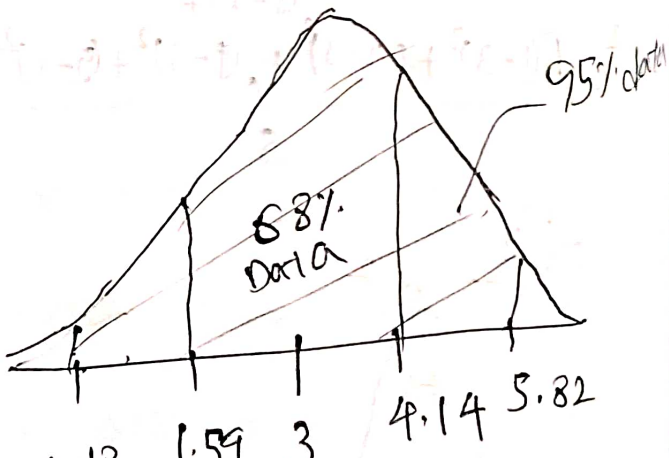
This is the square root of variance.

e.g

$\{1, 2, 3, 4, 5\}$

$\mu = 3, \sigma^2 = 2$

$\sqrt{2} = 1.41$



with in

$\pm 1\sigma$ (67% data)

$\pm 2\sigma$ (95% data)

$\pm 3\sigma$ (99.7% data)

④ Percentile & Quartile

percentages = {1, 2, 3, 4, 5, 6, 7, 8}.

percentage of even number

$$= \frac{\text{No. of evens}}{\text{Total no. of Numbers}} = \frac{4}{8} = 0.5 = 50\%$$

Percentile

Defⁿ: A percentile is a value below which a certain percentage of observation lie.

99 percentile = It means the person has got better marks then 99% of the other students.

Dataset = {2, 2, 3, 3, 5, 5, 5, 6, 7, 8, 8, 8, 8, 8, 9, 9, 10, 11, 11, 12}

What is the percentile rank of 10?

$$\text{Percentile rank of } X = \frac{\#\text{No. of value below } X}{n}$$

$$= \frac{16}{20} = 80 \text{ percentile.}$$

$$\text{Percentile rank of } 6 = \frac{7}{20} = 35 \text{ percentile}$$

Q. Find 25 percentile

$$\text{Value}_{(i)} = \frac{\text{Percentile}}{100} \times n$$

$$= \frac{25}{100} \times 20 = 5^{th} \text{ index}$$

Ans. = 5

Q. Find 30 percentile.

$$\text{value}_{(i)} = \frac{30}{100} \times 20 = 6^{th} \text{ index}$$

Ans = 5

56 percentile

$$\text{value}_{(i)} = \frac{56}{100} \times 20 = 11.2$$

Its not a whole number so

$$i = 11.2 + 1 = 12.2$$

and i = whole no. person is 12

Ans = 8

(5) 5 number Summery

(1) Minimum
(2) First Quartile (25 percentile) (Q1)
(3) median
(4) Third Quartile (75 percentile) (Q3)
(5) maximum.

} Remove the outliers (BOX plot)

{-15, 1, 2, 3, 3, 4, 5, 5, 5, 6, 6, 7, 8, 9, 9, 27}

Fence = [lower fence ↔ higher fence]

Lower Fence = Q1 − 1.5(IQR)

IQR = (Q3 − Q1)

Inter quartile range

Higher fence = Q3 + 1.5(IQR)

$Q_1 = \frac{25}{100} \times 16 = 4^{th}$ index, $Q_3 = \frac{75}{100} \times 16$

$= 12$ index

IQR = 7 − 3 = 4

LF = Q1 − 1.5(4)

= 3 − 6 = −3

HF = Q3 + 1.5(4)

= 7 + 6 = 13

[−3, 13]

outliers = −15, 27

① Minimum = -15

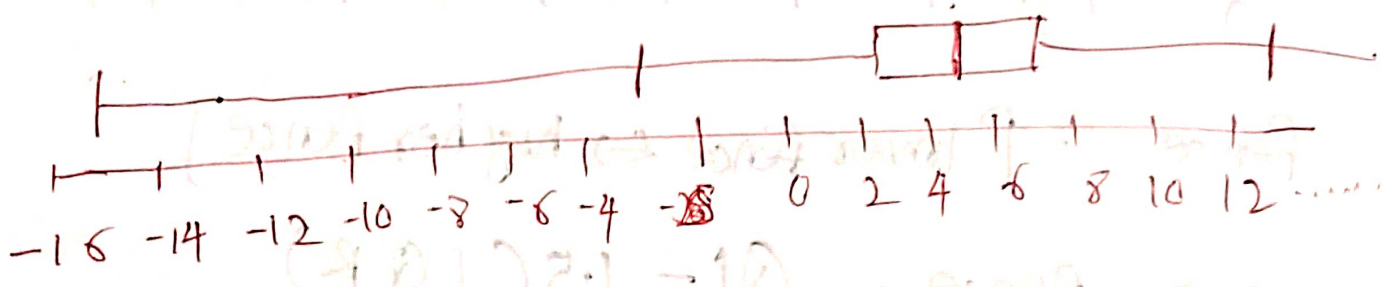
② Q1 = 3

③ Median = 5

④ Q3 = 7

⑤ Maximum = 27

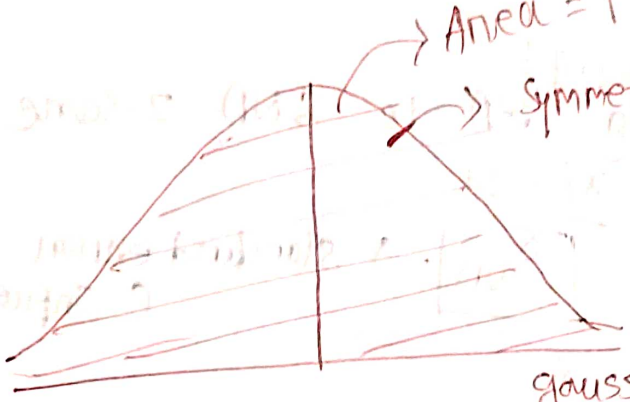
Box Plot



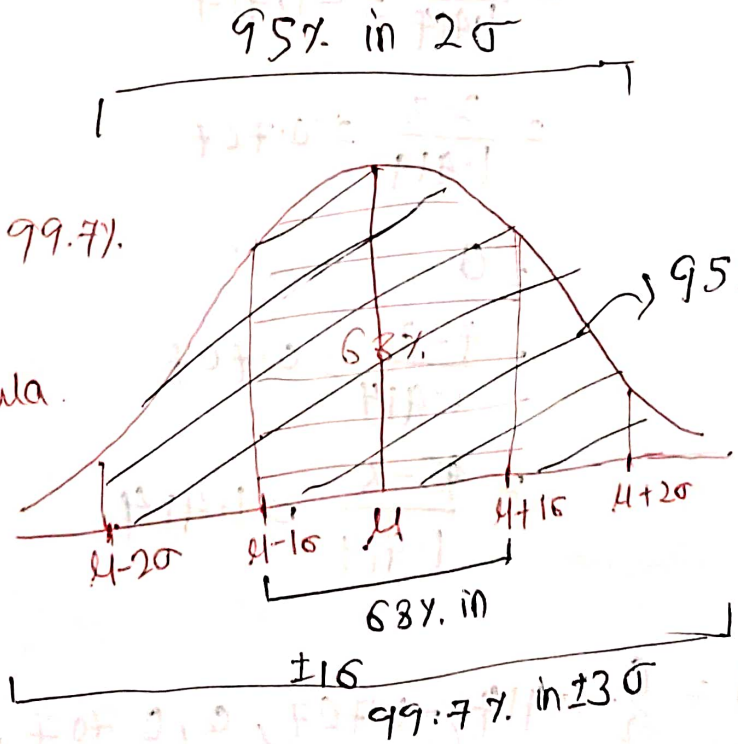
Agenda

1. Normal Distribution
2. Standard Normal Distribution
3. Z-score
4. Log Normal Distribution

1. Gaussian / Normal Distribution



Most of the data follows gaussian ~~symmetrical~~ distribution.



This 68-95-99.7% is called emperical formula.

How to decide a distribution is gaussian or not.?

Ans. Using QQ plot

② Standard Normal Distribution ($\mu=0, \sigma=1$)

$X \sim$ Gaussian Distribution (μ, σ)

↓ transform

$Y \sim$ Standard Normal Distribution (SND) ($\mu=0, \sigma=1$)

$X = \{1, 2, 3, 4, 5\}$

$\mu = 3, \sigma = 1.41$

for converting GD to SND z-score is calculated.

$$z\text{ score} = \frac{x_i - \mu}{\boxed{\sigma/\sqrt{n}}}$$ → standard error (Inferential stats)

↓

$$z\text{score} = \frac{x_i - \mu}{\sigma}$$ (∵ $n=1$ each data points are samples)

$$= \frac{1-3}{1.414} = -1.414$$

$$= \frac{2-3}{1.414} = 0.707$$

$$= 0$$

$$= \frac{4-3}{1.414} = 0.707$$

$$= \frac{5-3}{1.414} = 1.414$$

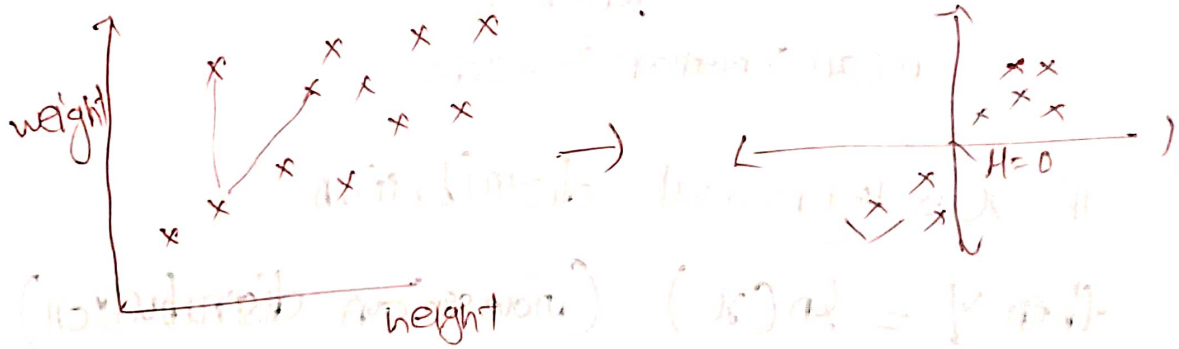
$Y = \{-1.414, -0.707, 0, 0.707, 1.414\}$

Why the above things required ?

why ?

Age (years)	weight (kg)	hight (cm)
24	72	150
26	78	160
32	84	165
33	92	170
34	87	150
35	83	180
38	80	175

→ To bring above data to same scale (standardization)



Normalization

[0, 5], [-1, 6], [3 - 9] (we provide range)

eg Min-max scaler

$$x_{scaled} = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

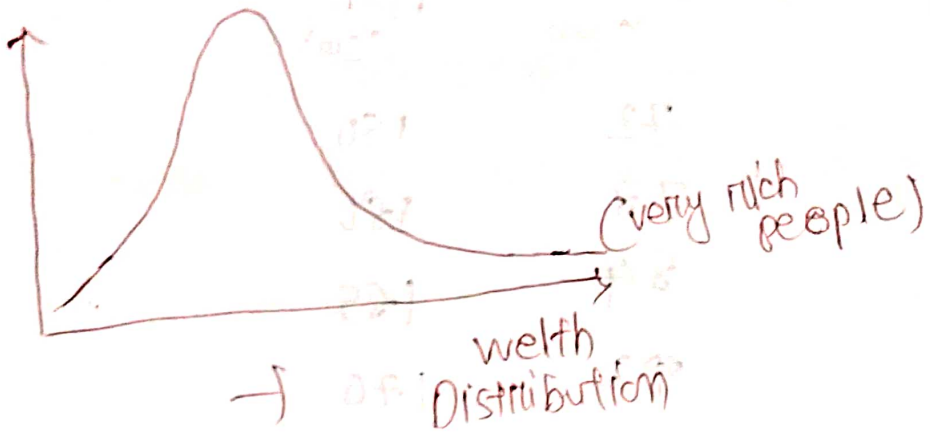
data will be ranged from (0, 1)

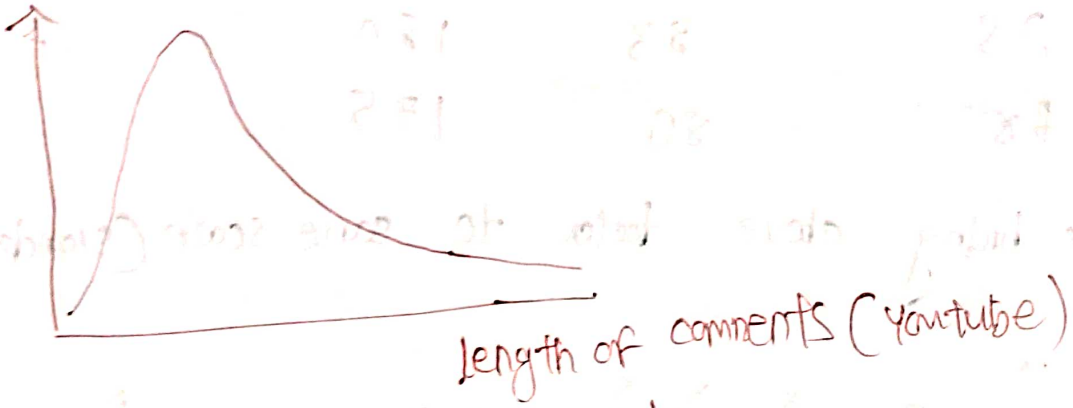
Standardization

Data
↓ standardize
Standard data
(H=0, σ=1)

[-3, 3]

④ Log Normal Distribution (LND)

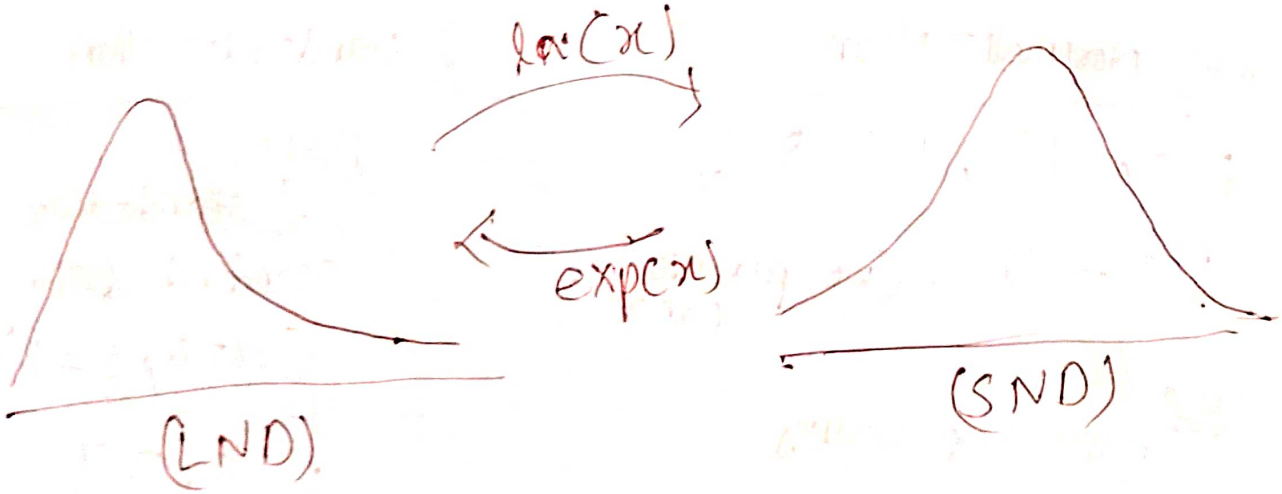




mean > median > mode

if $x \simeq$ log normal distribution

then $Y = \ln(x)$ (Gaussian distribution)



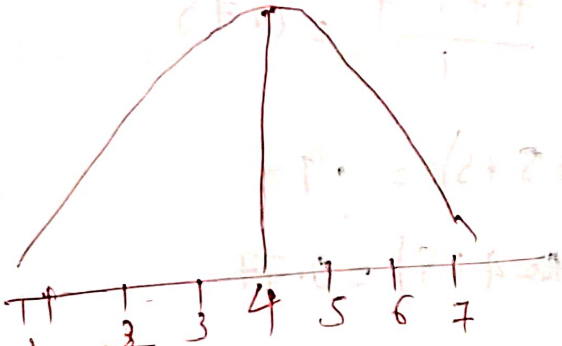
Finding % of area using Z score & Z table

Let

data = {1, 2, 3, 4, 5, 6, 7}

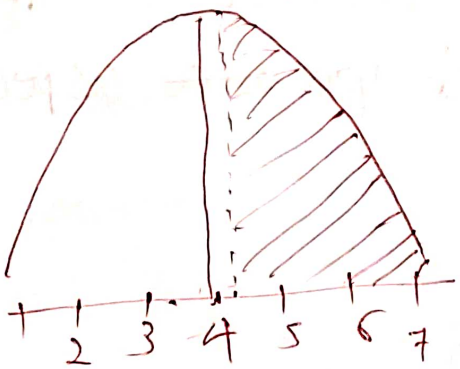
Let $\mu = 4$

$\sigma = 1$



$$\text{Zscore } z = \frac{x_i - \mu}{\sigma}$$

Find % Area ~~at 0.25~~ that falls above 4.25?



$$\text{Zscore} = \frac{4.25 - 4}{1} = 0.25$$

from Z table

Area = 0.59 (for below 4.25)

for above = $1 - 0.59 = 0.41 \Rightarrow 41\%$

Find area of curve between 4.75 and 5.75

$$Z\text{ score}(5.75) = \frac{5.75-4}{1} = 1.75$$

$$Z\text{ Score}(4.75) = \frac{4.75-4}{1} = 0.75$$

Area below for (5.75) = .96

Area below for (4.75) = 0.77

So between 4.75 & 5.75 = 0.96 − 0.77 = .19

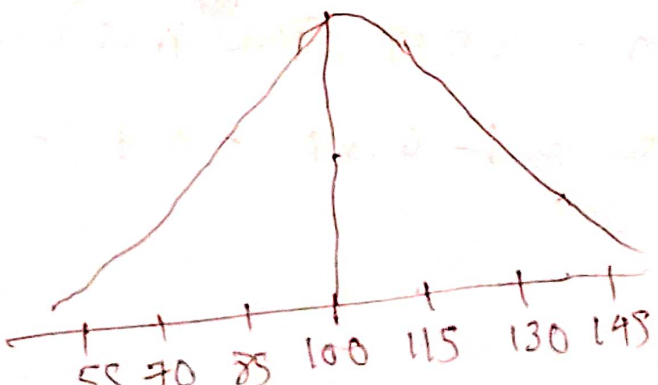
=) 19%

Q. In India the averge IQ is 100 with a standard deviation of 15. what is the percentage of population would you ~~exee~~ aspect to have an IQ

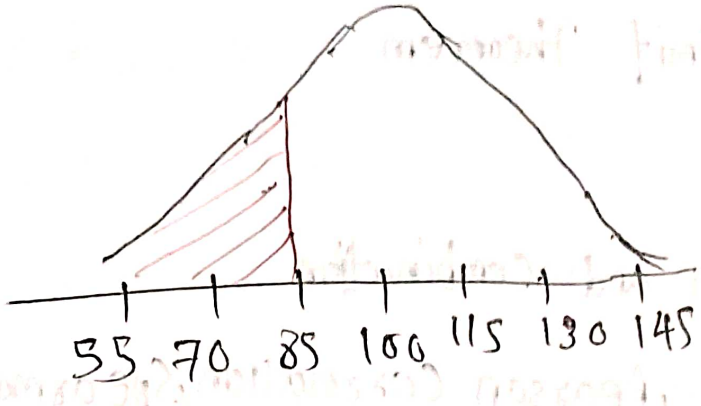
(i) lower then 85

(ii) Higher then 85

(iii) Between 85 and 100



(i)



$$Z\text{score} = \frac{85-100}{15} = -1$$

Area % = 0.1587 ⇒ 15.87%.

(ii) Higher then 85 = 1 − 0.1587

= 0.8413 ⇒ 84.13%.

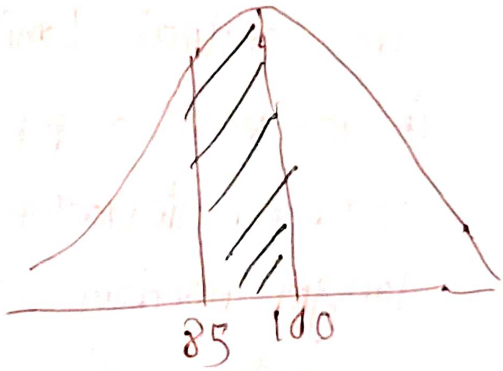
(iii) Between 85 and 100

Zscore(<85) = −1

Area = 0.1587

Zscore(<100) = 0

Area = 0.5



Area between 85 & 100 = 0.5 − 0.1587

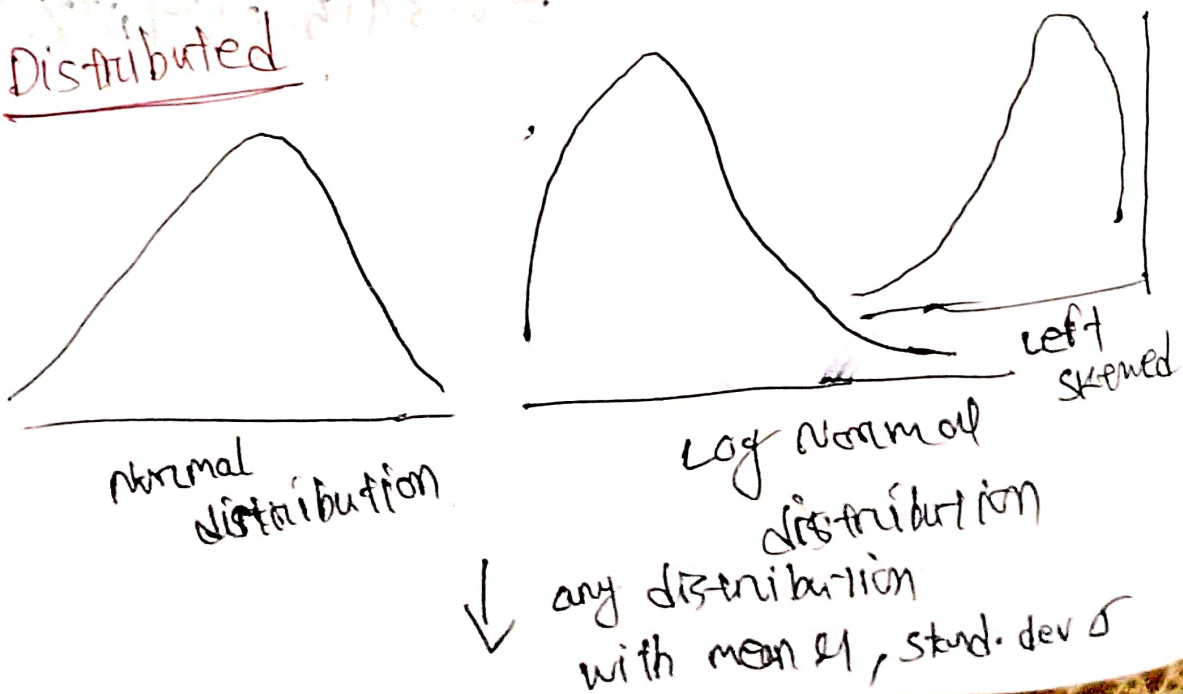
= 0.3413 ⇒ 34.13%.

Agenda

1. Central Limit Theorem
2. Probability
3. Permutation and combination
4. Covariance, Pearson correlation, Spearman correlation
5. Bernoulli's Distribution
6. Binomial Distribution
7. Power Law (Pareto Distribution)

1. Central Limit Theorem

The central Limit Theorem states that if you have a population with mean μ and standard deviation σ and take sufficiently large random sample ($n \geq 30$) from the population with replacement, then the distribution of the means will be approximately Normally Distributed



$S_1 = \{x_1, x_2, x_3 \ldots\ldots, x_n\}$

$S_2 = \{x_1, x_2, x_3 \ldots\ldots, x_n\}$

$S_3 = \{x_1, x_2, x_3, \ldots\ldots, x_n\}$

$\vdots$

$S_m = \{x_1, x_2, x_3, \ldots\ldots x_n\}$

The mean of of this m sample will follow gaussian/normal distribution, if no. of sample size > 30.

② Probability

Probability is the measure of likelihood of an event.

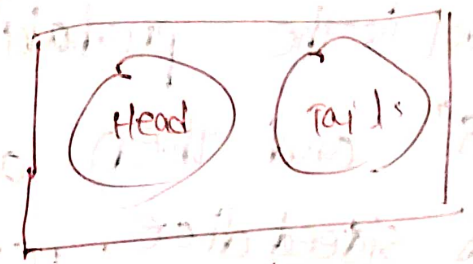
Eg. Tossing a fair coin $P(H) = 0.5$, $P(T) = 0.5$

Rolling a dice $P(1)$ or $P(2) \ldots P(6) = \frac{1}{6}$

(i) Mutual Exclusive event.

Two Events are mutually exclusive if they cannot occur at the same time.

Tossing a coin



(ii) Non-mutual exclusive event.

Two events occur at the same time.

Eg. A bag of marval containing Red, green yellow marbals pick two marbal randomly.

Deck of cards

→ what is the probability of choosing heart ♡ or Queen card ?

$$P(♡ \text{ or } Q) = P(♡) + P(Q) - P(♡ \text{ and } Q)$$

$$= \frac{13}{52} + \frac{4}{52} - \frac{1}{52} = \frac{16}{52}$$

Multiplication Rule

→ Dependent Events: Two events are dependent if they effect one another.

Bag of marbals { ○○○○ ○○○ }

$$P(W) = \frac{4}{7} \rightarrow P(R) = \frac{3}{6}$$

initial next

2nd effects due to event (1)

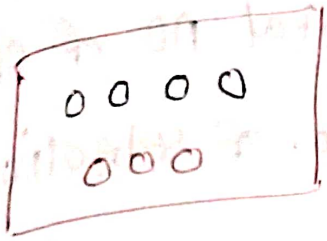
Q. what is the probability of rolling a "5" and then a "3" with a normal 6 sideed dice. $P(3) = \frac{1}{6}$, $P(5) = \frac{1}{6}$

Multiplication rule for Independent events

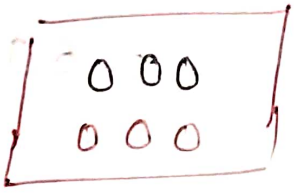
$$P(A \text{ and } B) = P(A) * P(B)$$

$$= \frac{1}{6} * \frac{1}{6} = \frac{1}{36}$$

(2) conditional probability (Naive Bayes is derived)



$P(O) = \frac{4}{7}$ → $P(Y/O)$



→ $\frac{3}{6} = \frac{1}{2}$

$P(O \text{ and } Y) = P(O) \times P(Y/O)$

$= \frac{4}{7} \times \frac{1}{2} = \frac{2}{7}$

→ Permutation

For a school children

{ Dairy milk, Kitkat, Milky bar, Snickers, 5 star } chocolates distributed in a order.

$\frac{5}{} \times \frac{4}{} \times \frac{3}{}$

$= 60$ ways. we can distribute in 60 ways

with permutation, order matters.

{ DM, KK, MB } } Both considered
{ KK, DM, MB } } All combinations.

Formula

$${}^{n}P_{r} = \frac{n!}{(n-r)!}$$

n = Total no. of object

r = No. of selection

$$= \frac{5!}{(5-3)!} = \frac{5 \times 4 \times 3 \times 2!}{2!}$$

$$= 60$$

~~e.g 2~~

Combination — Repetation will not occurs.

{DM KK MB}

X {KK DM MB} ←

Formula

$${}^{n}C_{r} = \frac{n!}{r!(n-r)!} = \frac{5!}{3! \times 2!} = \frac{5 \times 4 \times 3!}{3! \times 2}$$

$$= 10$$

④ Co-variance [feature selection]

Age	weight
12	40
13	45
15	48
17	60
18	52

Age↑ weight↑

Age↓ weight↓

$$cov(X,Y) = \sum \frac{(x_i - \bar{x}) * (y_i - \bar{y})}{n-1}$$

Hence

$\bar{X} = 15$, $\bar{Y} = 51$,

$$cov(X,Y) = \frac{(12-15)(40-51) + (13-15)(45-51) + (15-15)(48-51) + (17-15)(60-51) + (18-15)(62-51)}{5-1}$$

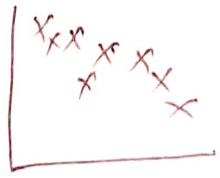
$$= \frac{33 + 12 + 18 + 33}{4}$$

$$= \frac{96}{4} = \boxed{24}$$ (+ve covariance)

$X\uparrow$ $Y\uparrow$

Co-variance

if) -Ve covariance $X\uparrow$ $Y\downarrow$



if) Zero covariance — NO co-relation

(5) Pearson co-relation coefficient. $(-1, 1)$

$$\rho(x,y) = \frac{cov(x,y)}{\sigma x \cdot \sigma y}$$

→ Restrict limit between $(-1,1)$

→ Tells how strengthen two more features are

→ Disadvantage: Useful for linear data.

Spearman's Co-relation

$$\gamma_s = \frac{COV(R(x), R(y))}{\sigma(R(x)) * \sigma(R(y))}$$

X	Y	R(X)	R(Y)
10	4	4	1
8	6	3	2
7	8	2	3
6	10	1	4

$\bar{X} = 2.5$ $\quad \bar{Y} = 2.5$

$$COV(R(x), R(y)) = \frac{(4-2.5)(1-2.5)+(3-2.5)(2-2.5)+(2-2.5)(3-2.5)+(1-2.5)(4-2.5)}{4-1}$$

$$= \frac{-2.55 - 0.25 - 0.25 - 2.55}{3}$$

$$= -1.87$$

$$\sigma(R(x)) = \frac{(4-2.5)^2 + (3-2.5)^2 + (2-2.5)^2 + (1-2.5)^2}{4-1}$$

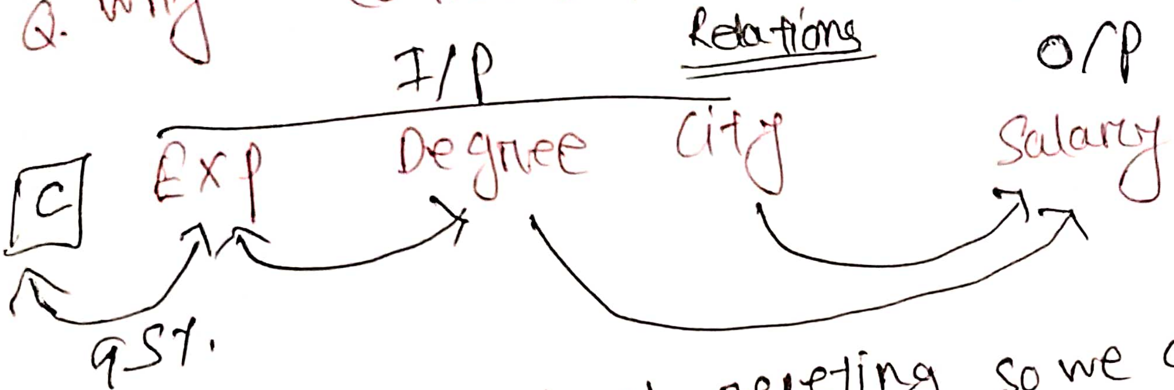
$$= \frac{2.55 + 0.25 + 0.25 + 2.55}{3}$$

$$= 1.87$$

$$\sigma(R(y)) = 1.87$$

$$r_s = \frac{-1.87}{1.87 \times 1.87} = -0.53$$

Q. why co-relation is used ?



C culums almost repeting so we can delete that column.