02 January 2018 21:14

Analytics → Inspecting → Stats
↓
Cleaning } Programming
↓
Transforming
↓
Modeling → Algorithms
↓
Decisions ⇒ Business

Ex Amazon Maps

Population → Sample ↳ Inferential Stats
↓
Descriptive stats

Describing Data

→ Central Tendency
→ Mean
→ Median
→ mode
→ Quartile
→ percentile

→ Variability
→ Variance
→ Standard deviation

→ Shape
→ Symmetry
→ Skewness
→ Kurtosis.

Variance & S.D

Mean → Avg
Median → Middle no. in Ascending Array

x_i
5 – x_1
9 – x_2
16 – x_3
17 – x_4
18 – x_5
65
$\mu = 13$

$(x_i - \mu) = 0$
–8
–4
3
4
5
0
$\sum(x_i - \mu) = 0$

$(x_i - \mu)^2$
64
16
9
16
25
130 → SSD
$\sum(x_i - \mu)^2$

$\sigma^2 = \dfrac{\sum(x_i - \mu)^2}{N}$

$\dfrac{130}{5} = 26$

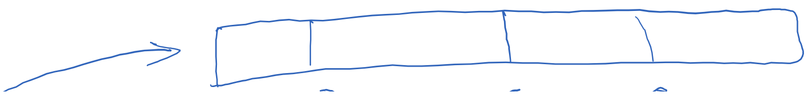
$SD = \sqrt{var}$
$SD = \sqrt{\sigma^2}$
$= \sigma$

Mode → Number having highest frequency → Unimodal, bimodal or Multi Modal

10 1 2 3 1 1 1 1 1 1 1 1 = $\dfrac{14}{11}$ = 1.23 → mean
1 1 1 1 1 1 1 1 1 1 2 3 10 = 2.4
= 1 = Median
= 1 = mode

Sensitive

Quartile → Dividing the data in to parts
3 → Tertile

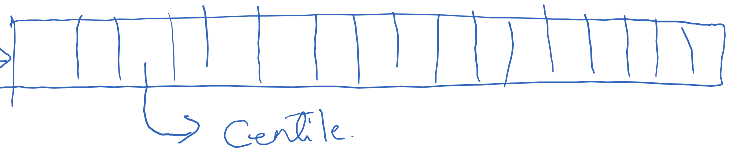


3 → Tertile

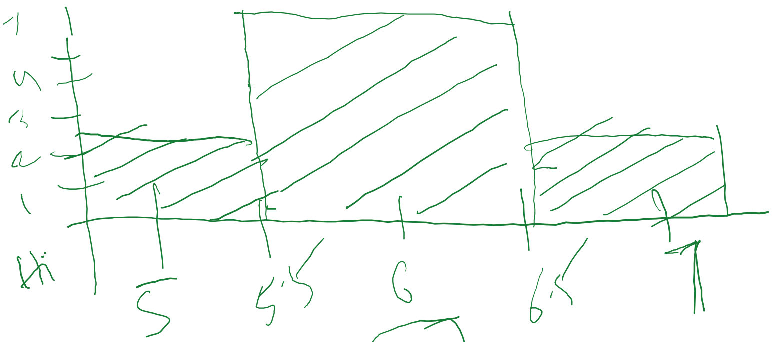
4 → Quartile

100 → Percentile

1 → Centile



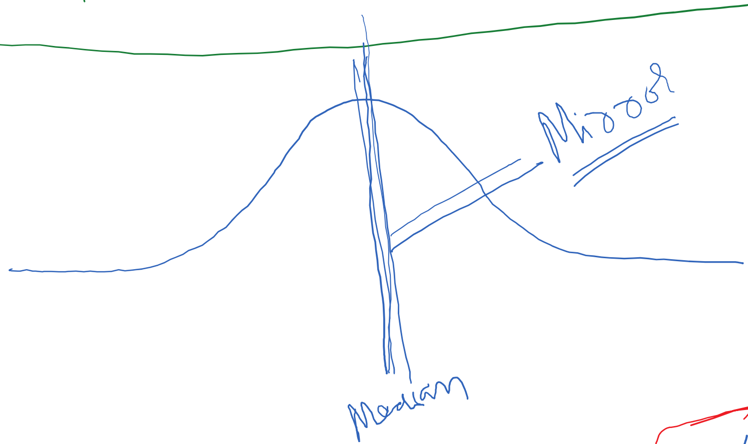
Shape

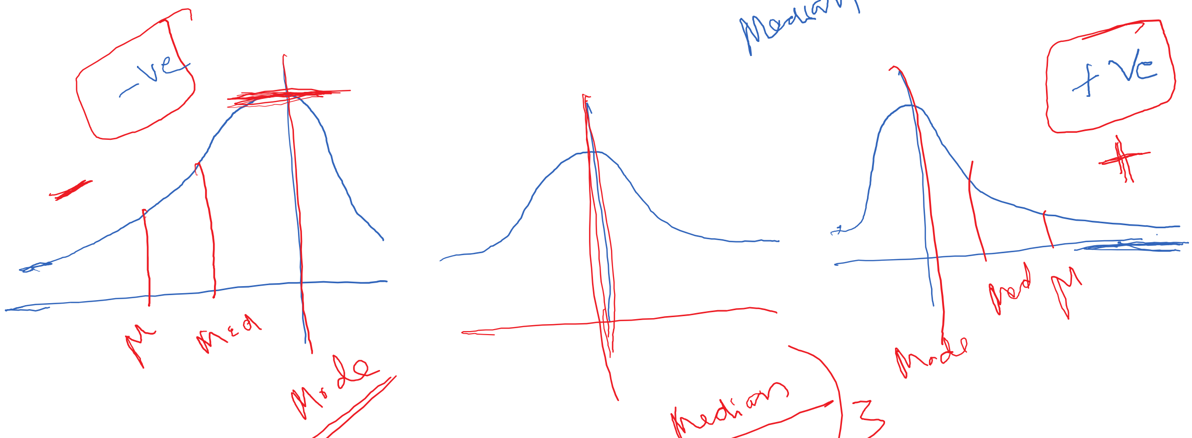


→ Symmetry →

X

Skewed





(mean − Median) 3 → Pearson

$$= \left(\frac{\text{Mean} - \text{Median}}{\text{SD}}\right) 3 \rightarrow \text{Pearson}$$

Peakedness

Nominal
Ordinal
Interval
Ratio

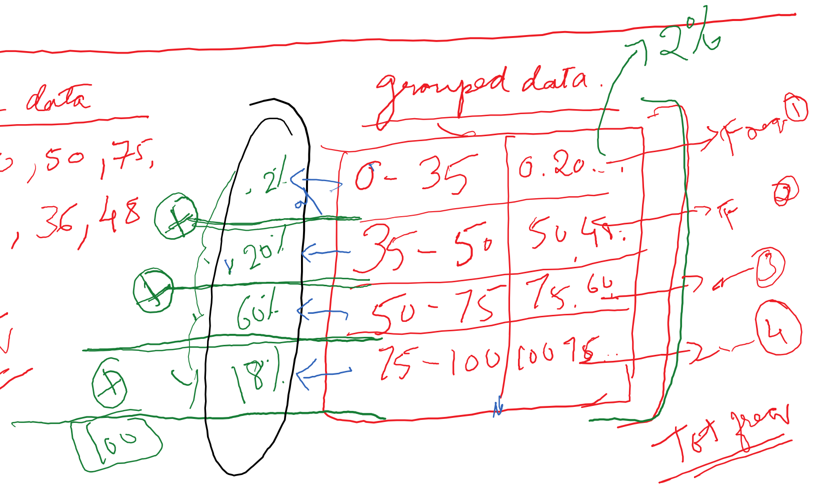
Qualitative &	Quantitative
Good, Bad	Age
Tall, short	Weight, height
complexion	IQ

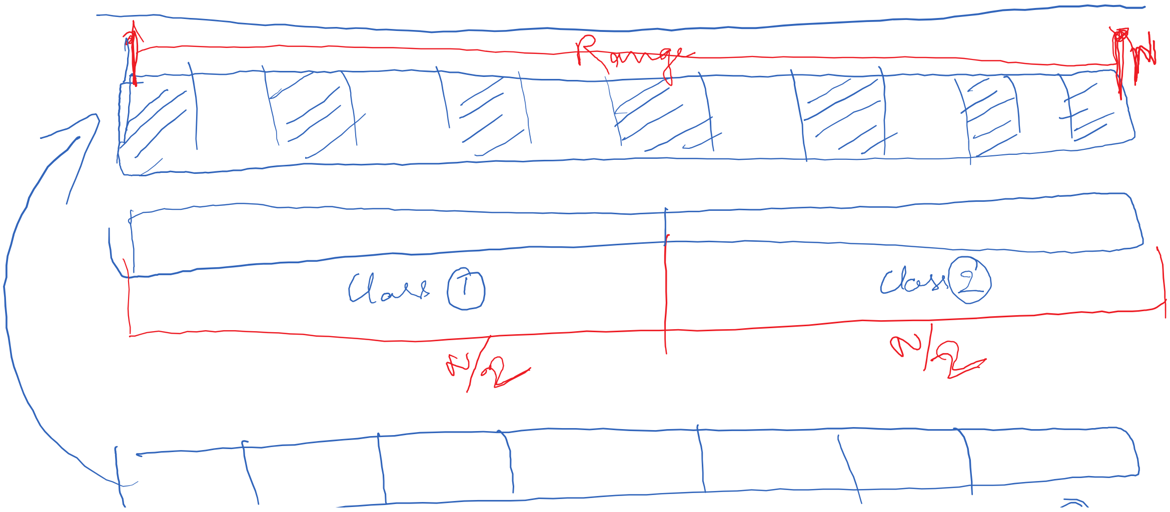
Ungrouped data

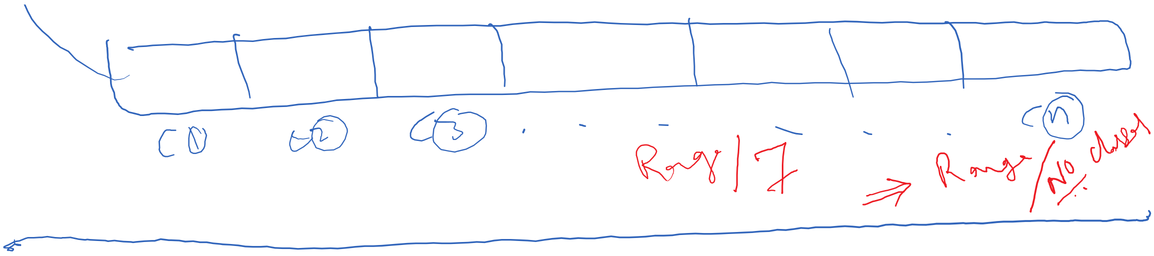
20, 30, 60, 50, 75,
90, 100, 0, 36, 48

Relative → Freq(1) / Total freq

grouped data

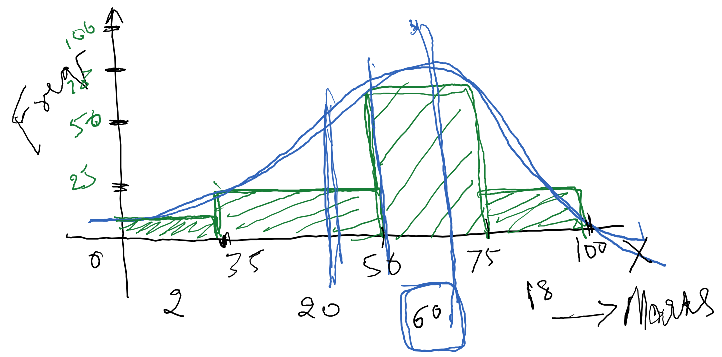




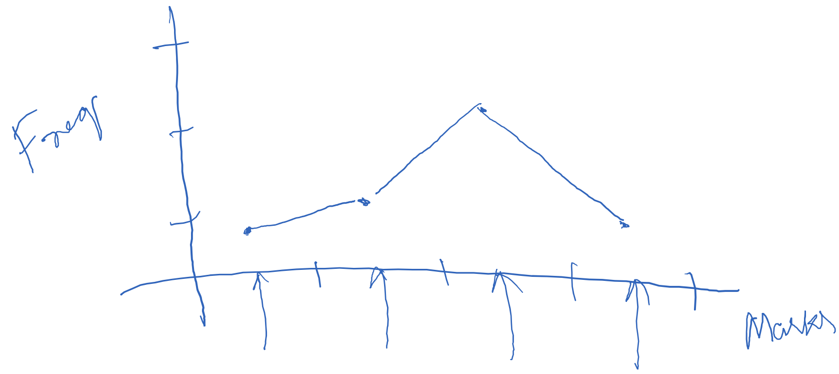


Quantitative graphs

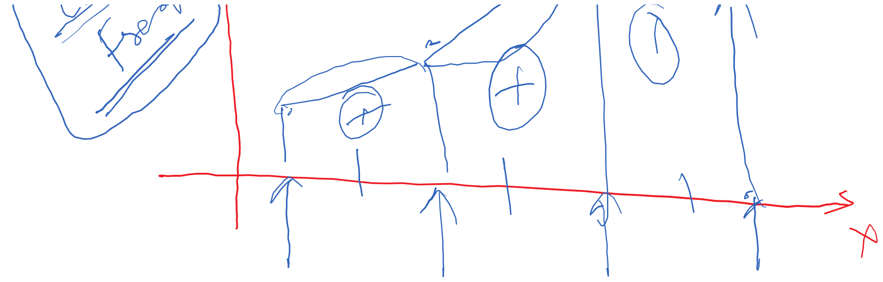
Histogram

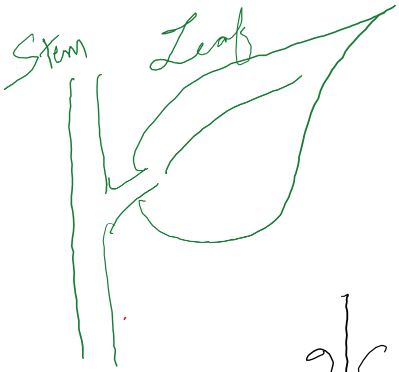


Freq polygon



Ogive





26, 34, 48, 23, 11, 16, 18, 50

Stem	Leaf
2	6 3
3	4
4	8
1	1 6 8
5	0

2 | 6

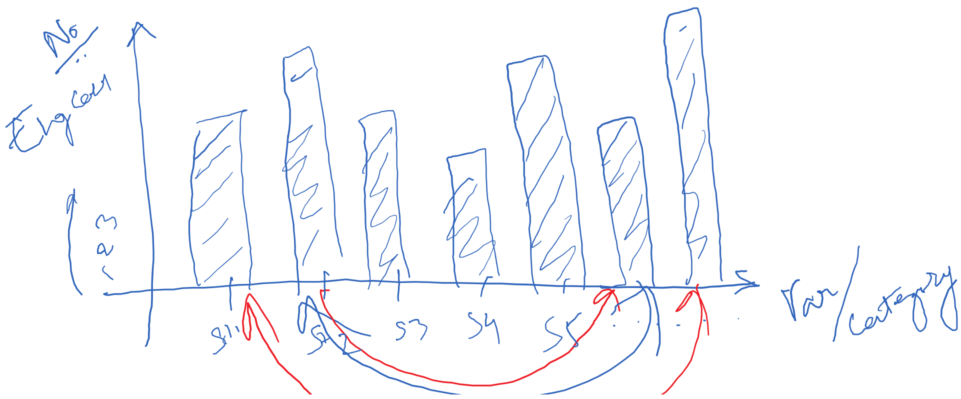
Stem ← → Leaf

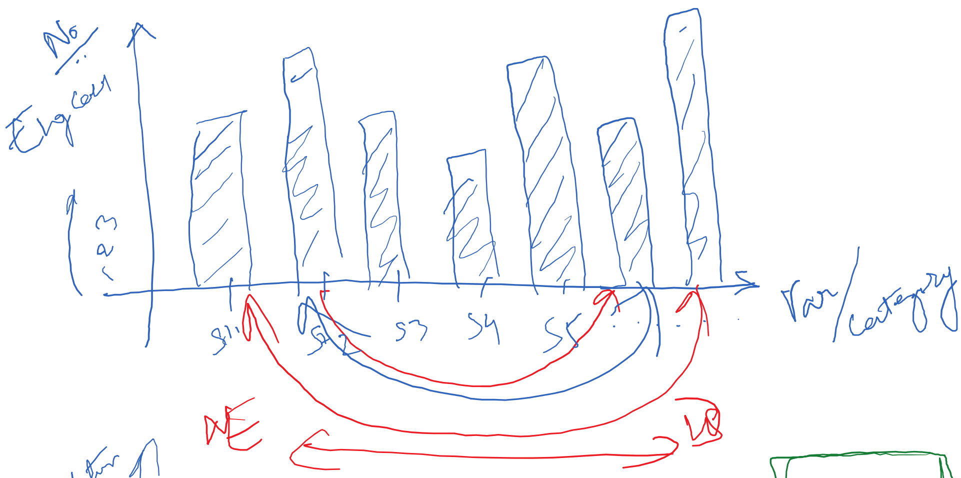
1	1 6 8
2	6 3
3	4
4	8
5	0

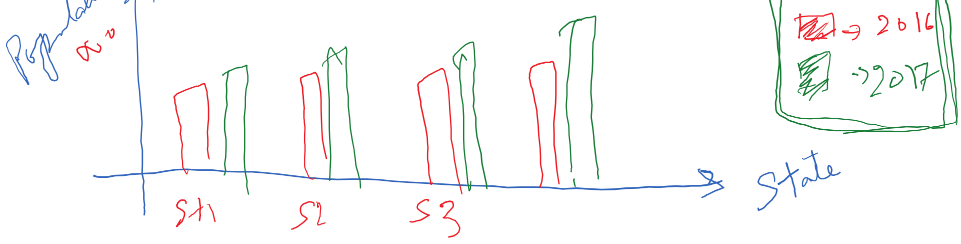
8				
6	3			
1	6	4	8	0
1	2	3	4	5

Qualitative graphs

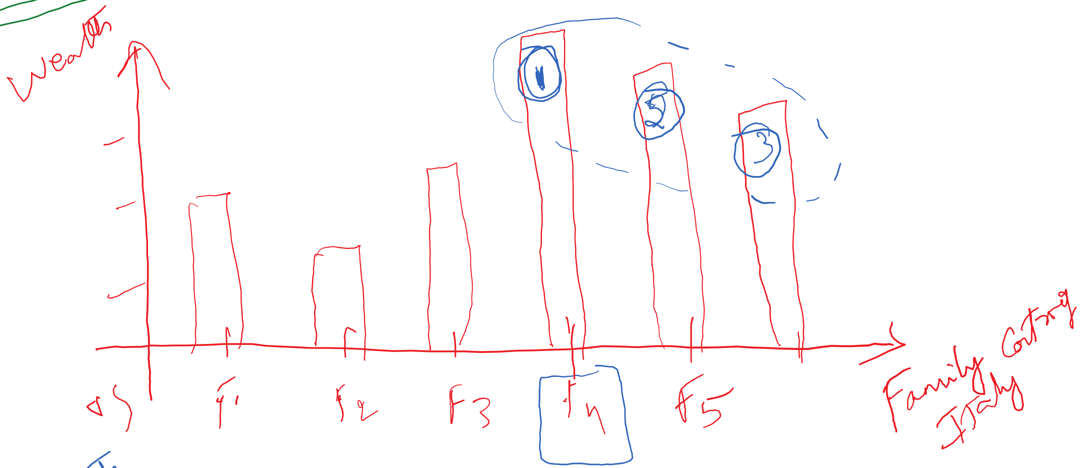
Barchart

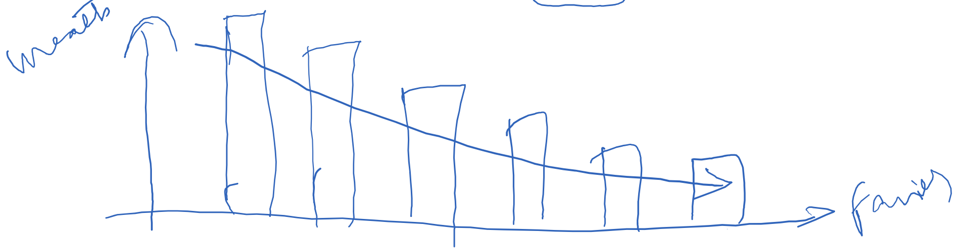


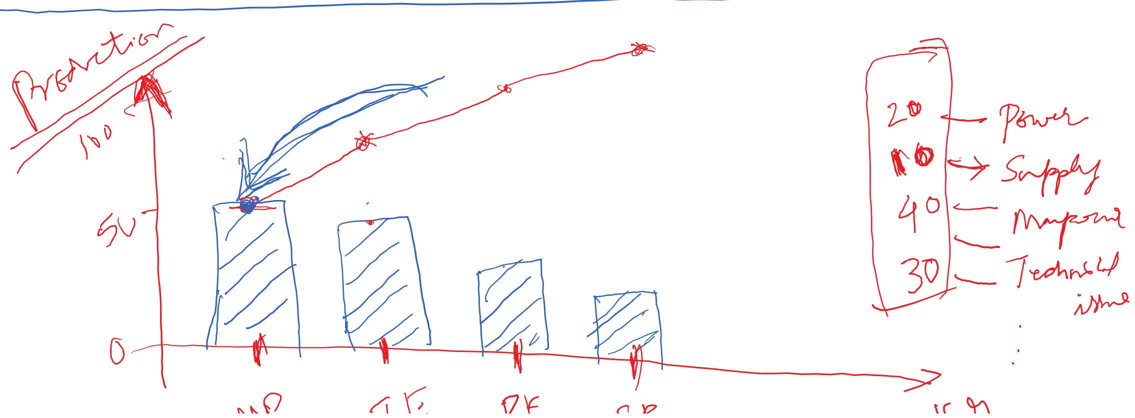


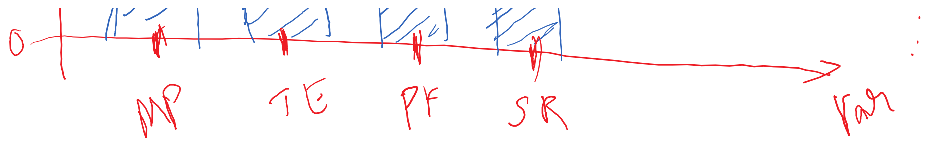


Pareto chart → Application of Bargraphs

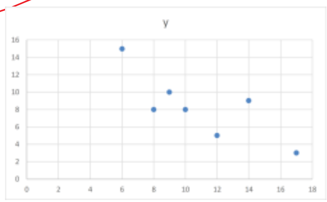


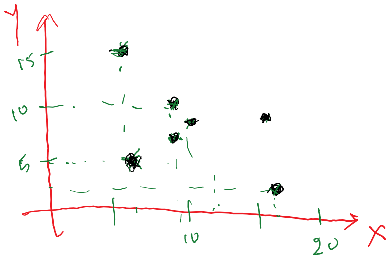






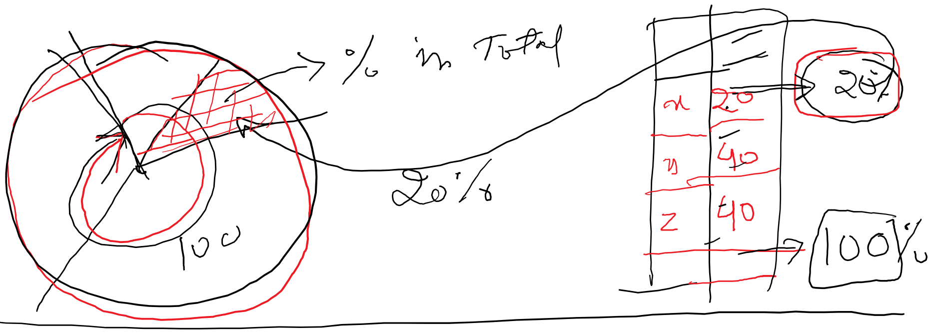
Scatter plot

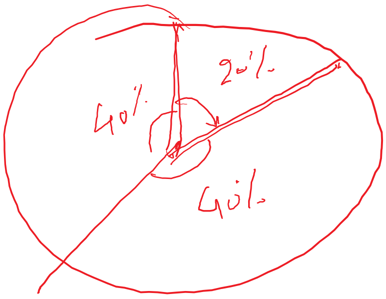




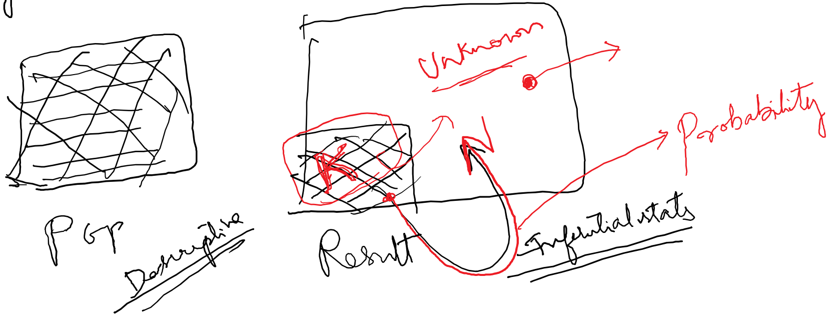
x	y
12	5
17	3
9	10
6	15
10	8
14	9
8	8

Pie chart





Probability



Experiment → Process to generate an outcome

Contests, Health checkup, Testing medicine

Event → The outcome of my experiment

Population of India

B.P/B.S/Hight
BMI

effectiveness

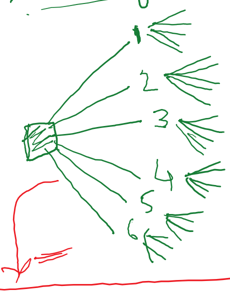
Event → "The outcome of any experiment"

Population of India

B.P / B.S / Hight / BMI

effectiveness

Elementary Events



Even no → Even → 2, 4, 6

H_2O → H, O

Elements

1,1	2,1	3,1	4,1	5,1	6,1
1,2	2,2	3,2	4,2	5,2	6,2
1,3	2,3	3,3	4,3	5,3	6,3
1,4	2,4	3,4	4,4	5,4	6,4
1,5	2,5	3,5	4,5	5,5	6,5
1,6	2,6	3,6	4,6	5,6	6,6

→ Sample space

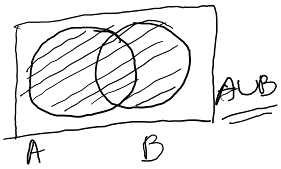
→ $1/36$ (4,4)

→ $6/36 = 1/6$

1 2 3 4 5 6 $\boxed{SP}$ $= \frac{1}{6}$

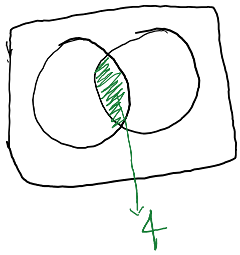
Union $A = \{1, \boxed{4}, 7, 9\}$ $B = \{2, 3, \boxed{4}, 5, 6\}$

$A \cup B = \{1, 2, 3, 4, 5, 6, 7, 9\}$



$A \cup B$

Intersection $A \cap B = \{4\}$



$A \cap B$

Trial

Mutually exclusive events

Collectively Exhaustive

Independent event

→ Diesel → Black → Automatic
→ Petrol → Blue → Manual
→ Red

$\boxed{m\ n \text{ Rule}}$

→ petrol → Blue → Manual
→ Red

2 3 2 = 12 options

Rule

⇒ 1 → 6 = 6^1

2 → 36 = 6^2

3 → 216 = 6^3

N = population size

n = sample size

→ $\boxed{N^n}$

Sampling population → $\boxed{\text{With Replacement}}$

1	2	3	4	5	6

1 → 5
2 → 5

Sampling population ⇒ without Replacement

1/14 1/16 1/15

N = 16
n = 3

$$\boxed{{}_N C_n = \frac{N!}{n!\,(N-n)!}} = \frac{16!}{3!\,(13)!} = 560$$

P(x)

Coin
Dice

P(x)

Marginal

X & Y → Independent ⇒

X or
Y or
X & Y

P(X ∪ Y)

Union

X & Y

P(X ∩ Y)

Joint probability

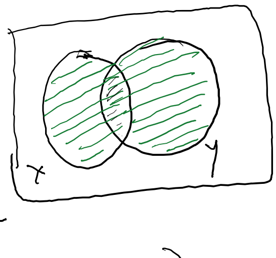
X given Y

P(X | Y)

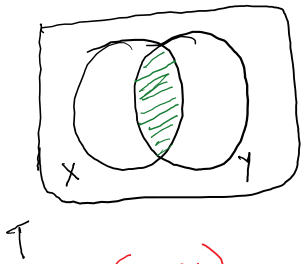
Conditional probability

Total possible prob

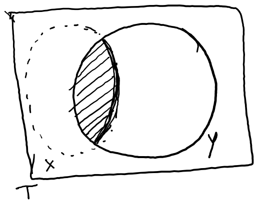
Total possible probs



Total possible probs



Sub group of population



$$P = \frac{\text{Sample}}{\text{population}}$$

$$P(X \cup Y) = P(X) + P(Y) - P(X \cap Y)$$

$$P(X \cap Y) = P(X) \cdot P(X|Y) = P(Y) \cdot P(Y/X)$$

$P(X \cup Y) = P(X) + P(Y) - P(X \cap Y)$ → Normal

$P(X \cup Y) = P(X) + P(Y) - 0$ → Mutually exclusive

	Gender M	F	
Manager	8	3	11
Professional	31	$\boxed{13}$	44
Technical	52	17	69
Clerical	9	22	31
	100	55	155

$$P(\text{Female} \cup \text{Professional}) \quad (X, Y)$$

$$= P(X) + P(Y) - P(X \cap Y)$$

$$= \frac{55}{155} + \frac{44}{155} - \frac{13}{155}$$

$$= 0.35 + 0.28 - 0.08$$

$$\boxed{= 0.55}$$

→ Cross table / Contingency table

	Gender M	F	
Manager	8	3	11
Professional	31	$\boxed{13}$	44
Technical	52	17	69
Clerical	9	22	$\boxed{31}$

Joint prob table

	M	F	
M	0.05	0.019	$\boxed{0.07}$
P	0.2	$\boxed{0.08}$	$\boxed{0.28}$
T	0.33	0.109	0.44

Technical			
Clerical	9	22	31
	100	55	155

T	0.33	0.10	0.4
C	0.05	0.14	0.2
	0.64	0.35	1

P(Female ∪ Professional)

$P(F) + P(P) - P(F \cap P) = \boxed{0.55}$

$0.35 + 0.28 - 0.08$

$P(X \cap Y) =$

T

$\boxed{P(X \cap Y) = P(Y) \cdot P(X|Y), \; P(X) \cdot P(Y|X)}$

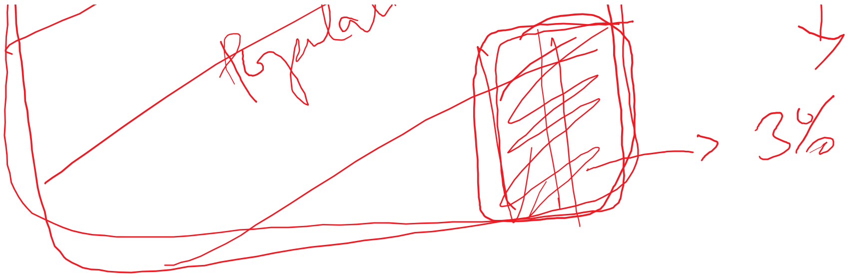
X & Y are independent

$\boxed{P(X \cap Y) = P(X) \cdot P(Y)}$ $\quad$ 0.3 $\quad$ 0.7

→ 3% defective → 2 production defective

$P(d) = \frac{3}{100} = 0.03$

$\boxed{0.009}$ $\quad 0.009 \times 0.03 = \quad 27$



Conditional prob $= P(X|Y) = \dfrac{P(X \cap Y)}{P(Y)}$

< Bayes Theorem > → change in the Input values

$\boxed{Y \;|\; X}$

Special Case

$$P(X|Y) = \frac{P(X\cap Y)}{P(Y)} = \frac{P(X)\,\boxed{P(Y|X)}}{\cancel{P(Y)}} \qquad \cancel{P(Y)}$$

$$= P(X)$$

	D	E	F	G	
A	0.12	0.05	0.04	0.07	0.28
B	0.15	0.03	0.11	0.06	0.35
C	0.14	0.09	0.06	0.08	0.37
	0.41	0.17	0.21	0.21	$\boxed{1}$

$P(B|F) = \dfrac{P(B\cap F)}{P(F)} = \dfrac{0.11}{0.21} = 0.5$

$P(G|C) = \dfrac{P(G\cap C)}{P(C)} = \dfrac{0.08}{0.37} = 0.21$

$P(D|F) = \dfrac{P(D\cap F)}{P(F)} = \dfrac{0}{\text{[illegible]}} = 0$

Discrete & Continuous Distribution

Discrete → Count

Continuous → Measure

→ Random Variable → Discrete RV

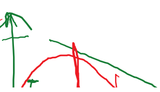
Random Variable → $\boxed{\text{Continuous RV}}$

Event	Probability
0	0.37
1	0.31

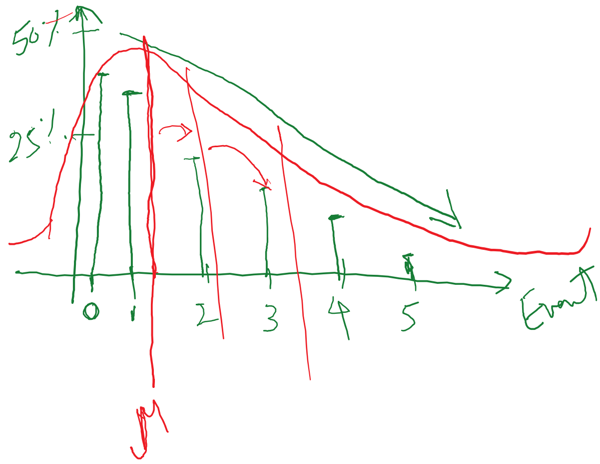
/ 8

(1)

50%



0	0.37
1	0.31
2	0.18
3	0.09
4	0.04
5	0.01



$$\mu = \sum \left(x \cdot P(x)\right)$$

$$\sigma^2 = \sum \left[(x-\mu)^2 \cdot P(x)\right] \qquad \sigma = \sqrt{var}$$

Binomial dist

$$\boxed{(p + q) = 1 \Rightarrow p = 1 - q}$$

$$\mu = x \cdot P(x) = \boxed{n \cdot p}$$

$\Rightarrow$ $\boxed{70\%}$ $\longrightarrow$ 4 ppl in Random (N) $\rightarrow$ 2 ppl (n) ✓ $\boxed{p = 0.7}$ $\boxed{q = 0.3}$

$$\boxed{0.26 = {}_4C_2\ 0.7^2 \cdot 0.3^2}$$

$$6 \times 2 \times 2$$

$$\boxed{{}_4C_2 \times p^2\ q^2}$$

$$\frac{4!}{2!\cdot 2!} = \frac{24}{4} = 6$$

$${}_4C_2$$

Difficult

~~Y Y Y Y~~
~~Y Y Y N~~
~~Y Y N Y~~ ×
Y Y N N ✓
~~Y N Y Y~~ ×
Y N Y N ✓
Y N N Y ✓
~~Y N N N~~
~~N Y Y Y~~
N Y Y N ✓
N Y N Y ✓
~~N Y N N~~
N N Y Y ✓
~~N N Y N~~
~~N N N Y~~
~~N N N N~~

6

0.0441

Y Y N N	$0.7 \times 0.7 \times 0.3 \times 0.3$
Y N Y N	$0.7 \times 0.3 \times 0.7 \times 0.3$
Y N N Y	

6

$$\boxed{P(x) = {}_NC_n \times p^n \times q^{N-n}}$$

$n = 2$, $N - n = 4 - 2 = 2$

$P(x) = {}^{n}C_{x} \, p^{x} \, q^{n-x}$

Poisson dist

→ No Trials

→ Independent

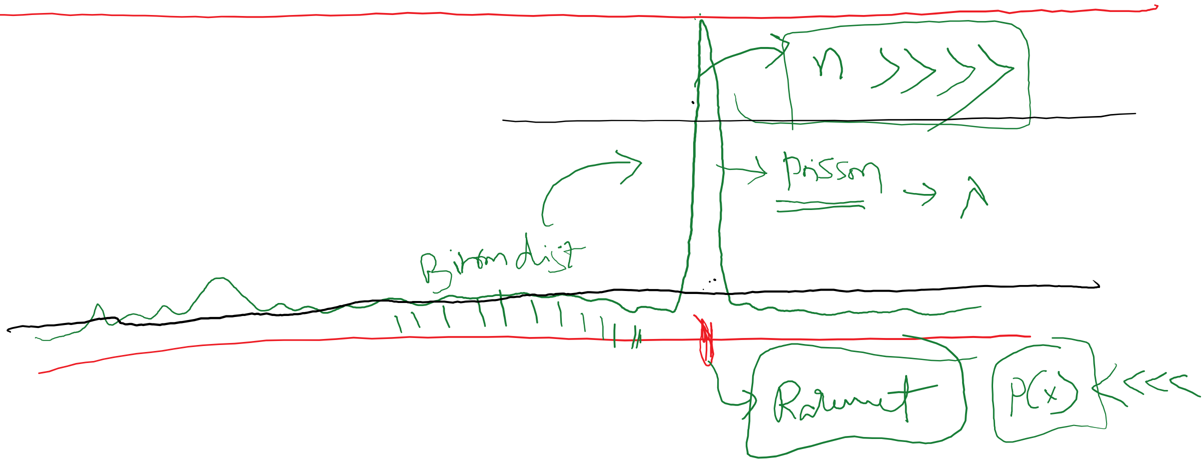
→ Rare → $P(x) <<<$ → Approx of Binomial

→ No events → >>>> 6 am ... 6 pm, time, $n >>>>$

λ → Mean → Variance → $S.D = \sqrt{\lambda}$

$$P(x) = \frac{\lambda^{x} \cdot e^{-\lambda}}{x!}$$

$e = 2.718$



Bank → 4 min = 3.2 → P(5) → 4 min

$$P(x) = \frac{\lambda^{x} e^{-\lambda}}{x!} \Rightarrow P(5) = \frac{(3.2)^{5} \cdot (2.71)^{-3.2}}{120} = \underline{\underline{0.115}}$$

$\lambda = 3.2$

(Excel for graph & example)

Hypergeometric Distribution

→ Discrete

→ Sample (S...) (Single %)

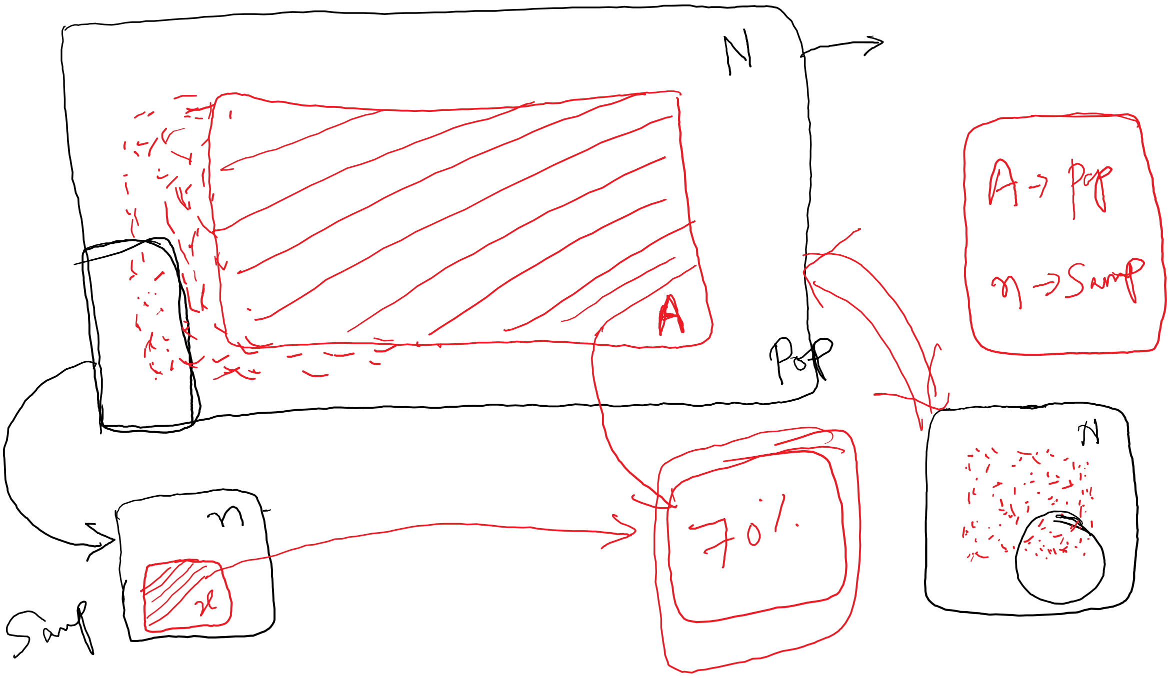
→ Discrete

→ Success / Failure (single o/p)

→ $\boxed{\text{Sampling without Replacement}}$

→ Population "N" is finite & Known

→ No of success in population is known "A"



$$P(x) = \frac{{}^{A}C_x \left({}^{N-A}C_{n-x}\right)}{{}^{N}C_n}$$

$$\mu = \frac{A \cdot n}{N}$$

$$\sigma^2 = \frac{A(N-A)\,n(N-n)}{N^2(N-1)}$$

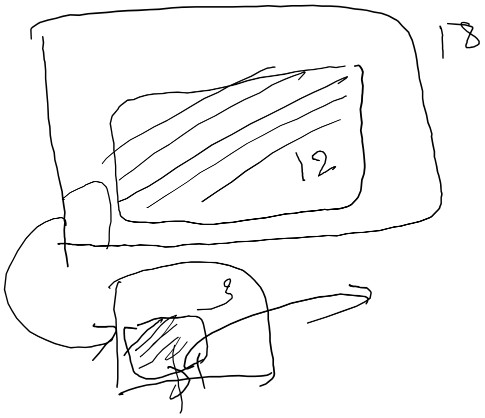
18 → Companies → N

12 → Silicon Valley → A

3 → n

1 ≥ → silicon valley

P(≥1) = ? → x

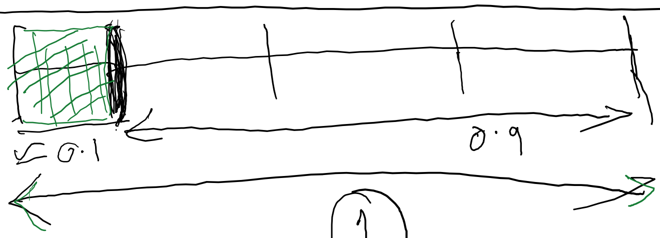


$\langle$ x = 1   x = 2   x = 3 $\rangle$

$x=1$ $\quad$ $x=2$ $\quad$ $x=3$

→ Ans + Ans + Ans →

(Cumulative)



Prob of a missile HIT – 97%
Miss – 3%

How many missing < 1/10L

① → 0.03 > $P(m) < 0.000001$

② → 0.009 >

③ → 0.000027 >

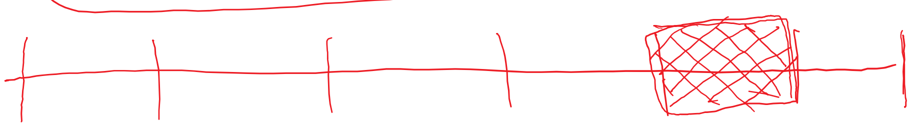
$\boxed{④ \to 0.00000081 < 1/10L}$

$${}^{n}C_0 \cdot (0.97)^0 \cdot (0.03)^n < 1/10L \quad ④$$

Geometric Dist → $\boxed{P(x) = p \cdot q^{x-1}}$

$p = 0.6$
$q = 0.4$

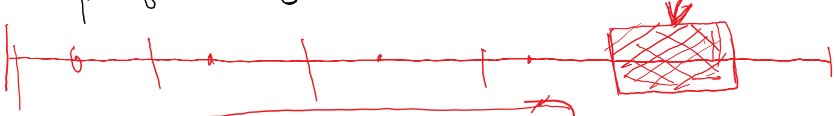
$\boxed{\text{Prob} \to 6 \text{ in the } 5^{th} \text{ ball}}$



$$P(5) = p \cdot q^{x-1} = (0.6) \times (0.4)^4 = \underline{0.015}$$

Negative binomial

p of 2^{nd} 6 in 5^{th} ball



$\boxed{P(x) = {}_{x-1}C_{r-1} \cdot p^r \cdot q^{x-r}} =$

$p = 0.6$
$q = 0.4$
$r = 2$
$x = 5$

$$\boxed{P(x) = {}_{n-1}C_{r-1} \cdot p^{r} \cdot q^{n-r}}$$

$r = 2$

$n = 5$

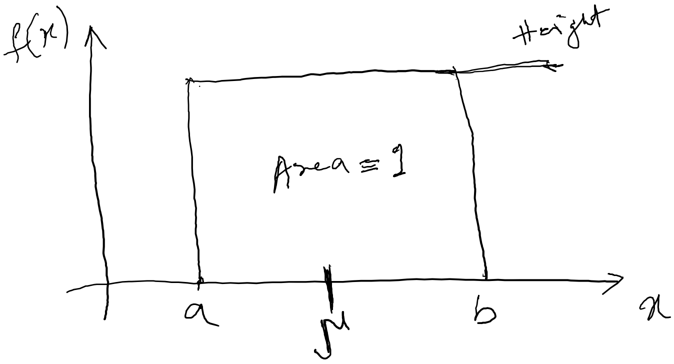
$${}^{4}C_{1} \cdot (0.6)^2 \cdot (0.4)^3 = 0.092$$

Continuous → Measured → Uniform, Normal (Z dist), Exponential

→ Area under the curve = 1 → $p(x)$

→ It can have ∞ values in a given Interval

Uniform



Area = $(b-a) \times$ Height

$1 = (b-a) \times$ Height

$$\boxed{\text{Height} = f(x) = \frac{1}{b-a}}$$

$$\boxed{\mu = \frac{a+b}{2}}$$

$$\boxed{SD = \sigma = \frac{b-a}{\sqrt{12}}}$$

$\mu = E(x)$

$\sigma^2 = E((x-\mu)^2)$

$= \boxed{E(x^2)} - \mu^2$

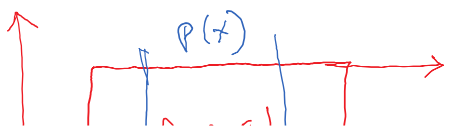
$$E(x^2) = \int_a^b \frac{x^2}{b-a} \cdot dx$$

$$\Rightarrow \frac{1}{3}\left(a^2 + ab + b^2\right)$$

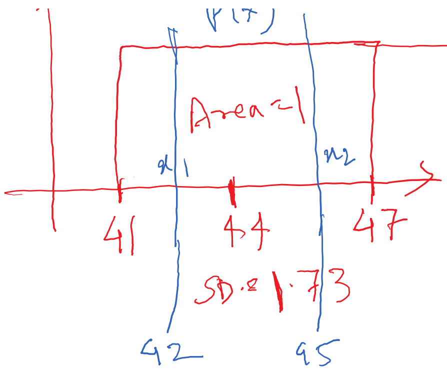
Derivation

$$\sigma^2 = \frac{1}{3}\left(a^2 + b^2 + ab\right) - \left(\frac{a-b}{2}\right)^2$$

$$\sigma^2 = \frac{1}{2}(b-a)^2 \Rightarrow \boxed{\sigma = \frac{b-a}{\sqrt{12}}}$$



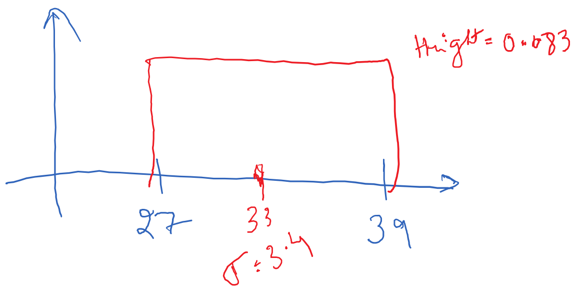
$f(x) = 0.166$



$$\boxed{P(x) = \frac{x_2 - x_1}{b - a}}$$

$a \le x_1 \le x_2 \le b$

$P(x) = \boxed{42 \,\&\, 45} = \frac{42-42}{47-41} = \frac{3}{6} = 0.5$



① Describe

$f(x) = 0.083$

$\mu = 33$

$S.D = 3.4$

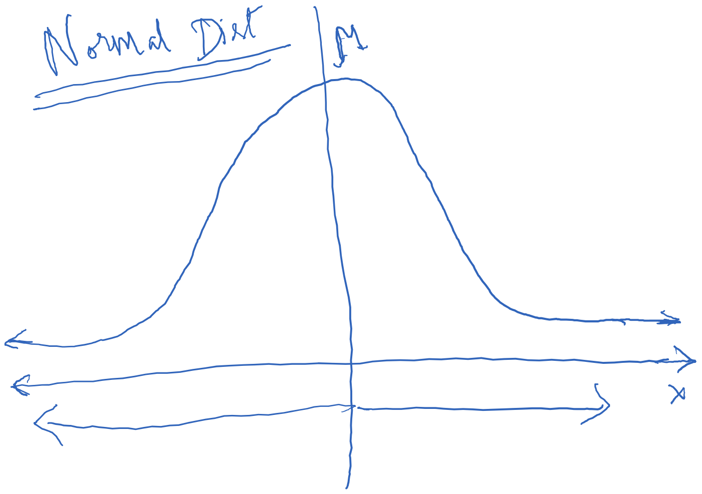
② Prob (30 to 35)

$P = \frac{35-30}{39-27} = 0.416$

③ prob (≤ 30)

$P = \frac{30-27}{39-27} = \frac{3}{12} = 0.25$

Normal Dist



→ Continuous dist

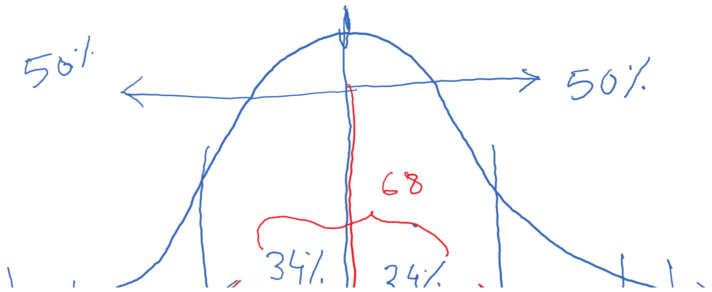
→ Symmetric → μ

→ Asymptotic to X axis $-\infty$ to $+\infty$

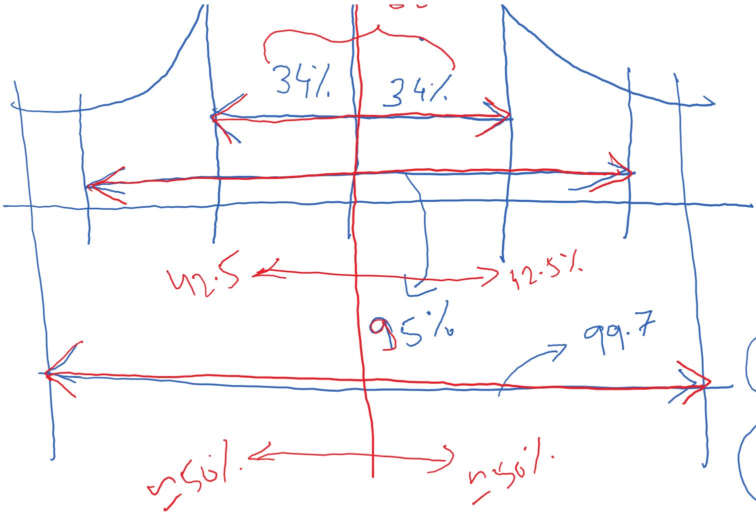
$\boxed{\mu = M_D = \text{Mode}}$

→ Area under curve = 1

→ Unimodal → single mode



$\boxed{\text{Chebishev's} \Rightarrow \text{Inequality}}$



⇒ Inequality

① $\boxed{68\% \cdot 1 \text{ SD}}$ from mean

34 34

② $\boxed{95\% \to 2 \text{ SD}}$ from mean

③ $\boxed{99.7\% \to 3\text{D}}$ from μ