* EDA and Probability Theory :-

① Mean : average , $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \dots + x_n}{n}$

② Standard deviation, $\sigma = \sqrt{\text{variance}} = \sqrt{\sigma^2}$

variance $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{(x_1-\mu)^2 + (x_2-\mu)^2 + \dots + (x_n-\mu)^2}{n}$

③ Median :- Given n observations $x_1', x_2', x_3', \dots, x_n'$

sort them $\Rightarrow x_1 < x_2 < x_3 < \dots < x_n$

n is even

Median $= \frac{\left(\frac{n}{2}\right)\text{term} + \left(\frac{n}{2}+1\right)\text{term}}{2}$

$x_1, x_2, x_3, x_4, x_5, x_6$

Median $= \frac{x_3 + x_4}{2}$

n is odd

Median $= \left(\frac{n+1}{2}\right)^{th}$ term

$x_1, x_2, \boxed{x_3}, x_4, x_5$

$\left(\frac{5+1}{2}\right) = 3^{rd}$ term

Median $= x_3$

④ Percentile , Quantiles, IQR

Percentile → divide given range in 100 parts

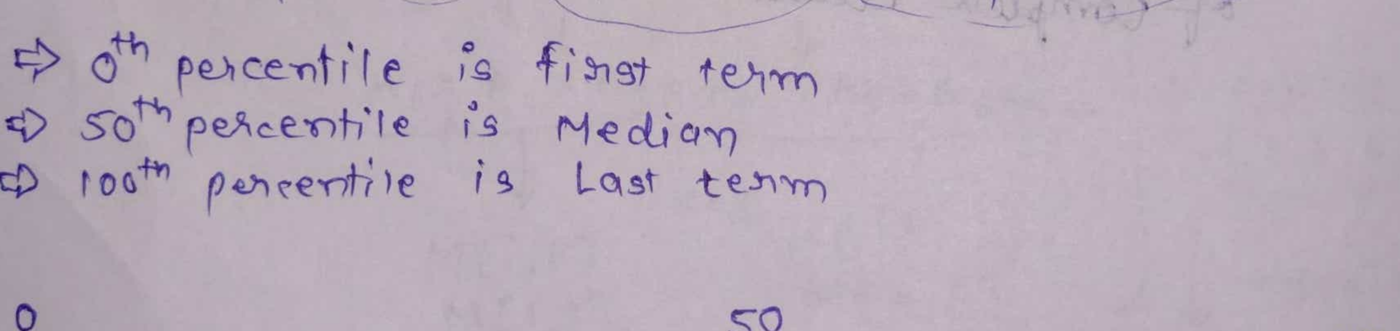
Quantiles → $25^{th}, 50^{th}, 75^{th}$ percentiles

IQR → 75^{th} percentile $- 25^{th}$ percentile

⇨ 0^{th} percentile is first term

⇨ 50^{th} percentile is Median

⇨ 100^{th} percentile is Last term



⇨ p^{th} percentile term $= \frac{n-1}{\left(\frac{100}{p}\right)} + 1 =$ int K + fraction f

p^{th} percentile $= x_K + f(x_{K+1} - x_K)$ (Linear interpolation)

$Q_1 = P(25) = 2.5$
$Q_2 = P(50) = \text{Median} = 5.5$
$Q_3 = P(75) = 7.5$

$\left.\begin{matrix} \\ \\ \end{matrix}\right\}$ Quantiles

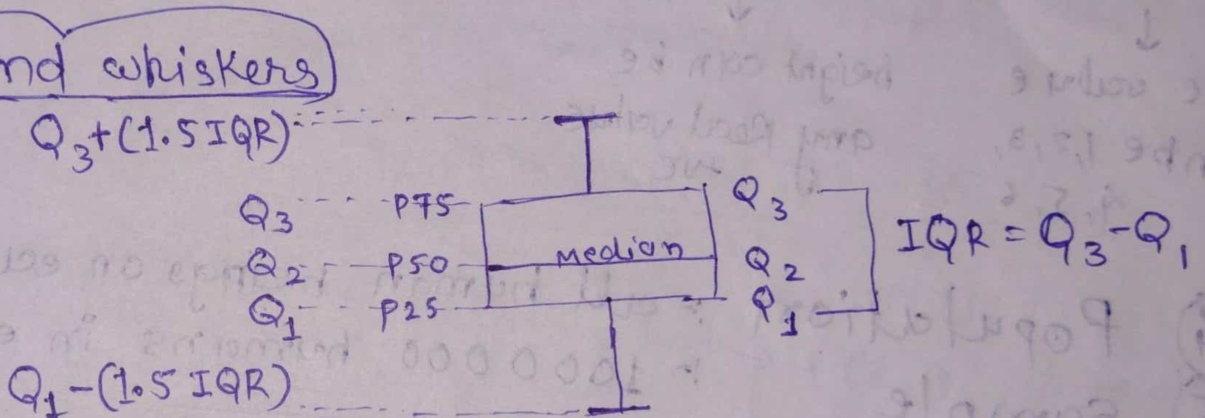
$\Rightarrow IQR = Q_3 - Q_1 = 7.5 - 2.5 = 5$

(5) MAD : Median Absolute deviation

Let M be the Median of $x_1, x_2, \ldots, x_n$

$$MAD = \text{Median}\left(\sum |x_i - M|\right)$$
$$= \text{Median}\left(|x_1 - M|, |x_2 - M|, \ldots, |x_n - M|\right)$$

* Boxplot and whiskers



50th per $= \dfrac{10-1}{\left(\frac{100}{50}\right)} + 1 = \dfrac{9}{2} + 1 = 5.5^{th}$ term $= 5^{th} + 0.5$

$\therefore P(50) = x_5 + 0.5(x_6 - x_5)$
$= 5 + 0.5(6-5)$
$= 5.5$

$\Rightarrow P(25) = \dfrac{10-1}{\left(100/25\right)} + 1 = \dfrac{9}{4} + 1 = 3.25 = 3^{rd} + 0.25$

$P(25) = x_3 + 0.25(x_4 - x_3)$
$= 3 + (0.25)(4-3)$
$= 3.25$

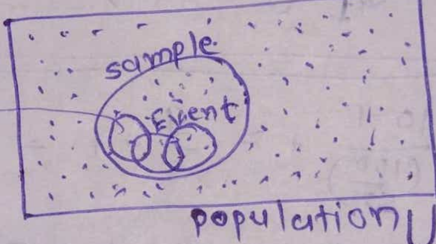
* **Probability** :-

1. Experiment → Roolling dice, Measure height, coin toss
2. Sample space → $\{1,2,3,4,5,6\}$, $R^+ - \{0\}$, $\{H, T\}$
   $\{1, 1.1, 2, 3, 26, 62.9, \ldots\}$
3. Random variable → X = Event → subset of S

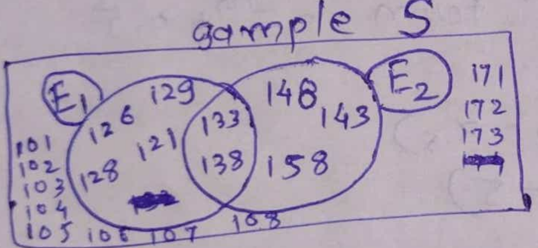
   getting dice value > 3
   getting height value > 125 cm
   getting H two times → $\{HH, HT, TH, TT\} = S$
   $X = \{HH\}$

Random variable:
- discrete → contable outcomes → dice value con be 1, 2, 3, 4, 5, 6
- continuous → un-contable (∞) outcomes → height can be any Real value +ve

4. Population :- all human beings on earth
5. Sample :- 1000000 humans in experiment



$E_i \subseteq S \subseteq U$



E_1 : height is between 120 to 140 , $\{120 \le x \le 140\}$
E_2 : height is between 130 to 160 , $\{130 \le x \le 160\}$

$E_1 = \{121, 129, 126, 128, 133, 138\}$ $n(E_1) = 6$
$E_2 = \{133, 148, 138, 158, 143\}$ $n(E_2) = 5$

Assume we have sample S of size 20 , $n(S) = 20 = n$

union $E_1 \cup E_2 = \{121, 129, 126, 128, 133, 138, 148, 158, 143\}$
$n(E_1 \cup E_2) = 9$

intersection $E_1 \cap E_2 = \{133, 138\}$
$n(E_1 \cap E_2) = 2$

$$n(E_1 \cup E_2) = n(E_1) + n(E_2) - n(E_1 \cap E_2)$$

$E_1^c = S - E_1 = \{101, 102, 103, 104, 105, 106, 107, 108, 148, 143, 158, 171, 172, 173\}$

$n(E_1^c) = n(S) - n(E_1) = 20 - 6 = 14$

→ Frequentist approach
1) Experiment is performed n times
2) $n(E)$: number of times event E occured

$\therefore P(E) = \lim_{n \to \infty} \frac{n(E)}{n}$ → In genral $P(E) = \frac{n(E)}{n(S)}$

classical $\left\{ = \frac{\text{favourable outcome}}{\text{total outcomes}} \right.$

$\Rightarrow P(E_1) = P(120 \le x \le 140) = \frac{n(E_1)}{n(S)} = \frac{6}{20} = 0.3 = 30\%$.

$\Rightarrow P(E_2) = P(130 \le x \le 160) = \frac{n(E_2)}{n(S)} = \frac{5}{20} = 0.25 = 25\%$.

$\Rightarrow P(E_1 \cup E_2) = P(120 \le x \le 160) = \frac{n(E_1 \cup E_2)}{n(S)} = \frac{9}{20} = 45\%$

$= P(E_1) + P(E_2) - P(E_1 \cap E_2)$ → $P(E_1 \cap E_2)$ is also written as $P(E_1 E_2)$

$= \frac{6}{20} + \frac{5}{20} - \frac{2}{20}$

$= 45\%$.

$\Rightarrow P(E_1^c) = 1 - P(E_1) = 1 - \frac{6}{20} = \frac{14}{20} = 70\%$

* $\boxed{P(E_1 \cup E_2) \le P(E_1) + P(E_2)} - P(E_1 \cap E_2)$ → ≥ 0 +ve value

* $P(E_1 \cup E_2 \cup E_3) = P(E_1) + P(E_2) + P(E_3) - P(E_1 E_2) - P(E_2 E_3) - P(E_1 E_3) + P(E_1 E_2 E_3)$

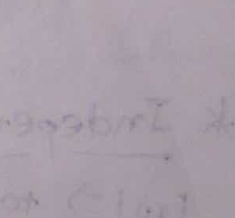
$\Rightarrow$ <u>AXIOMS OF PROBABILITY</u>

① $0 \le P(E_i) \le 1$

② $P(S) = 1$

③ if $E_1, E_2, \ldots, E_k$ are mutually exclusive events

$E_i \cap E_j = \phi$



then $P(E_1 \cup E_2 \cup E_3 \cup \ldots \cup E_k) = \sum_i P(E_i) = P(E_1) + P(E_2) + \ldots + P(E_k)$

* Conditional probability & Baye's theorem

$P(A|B) = \frac{P(A\cap B)}{P(B)}$

$P(B|A) = \frac{P(A\cap B)}{P(A)}$

$\rightarrow P(B|A)\cdot P(A)$

$$\boxed{P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}} \quad \neq 0$$

OR

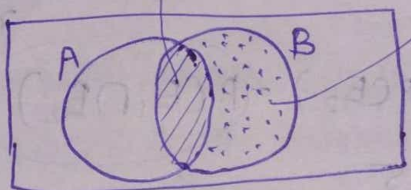
$$P(B|A) = \frac{P(A|B)\cdot P(B)}{P(A)} \quad \neq 0$$

$P(A|B) \neq P(B|A)$

B has already occured
Now what is $P(A)$?

Now $P(B) = P(B\cap A) + P(B\cap A^c)$

$(B\cap A)\cup(B\cap A^c)$
$= B\cap(A\cup A^c)$
$= B\cap U$
$= B$



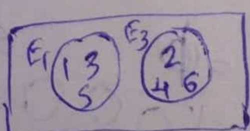
OR
$= P(B) + P(A\cap A^c)$
$= P(B) + 0$
$= P(B)$

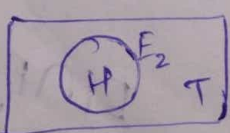
$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B\cap A) + P(B\cap A^c)}$$

$$P(A|B) = \frac{P(B|A)\cdot P(A)}{P(B|A)\cdot P(A) + P(B|A^c)\,P(A^c)}$$ → No need to know $P(B)$

* Independent event | Mutually exclusive event

Exp 1 => toss a coin
Exp 2 => throw a dice





E_1 and E_2 are independent
E_3 and E_2 are independent

$P(E_1\cap E_2) = P(E_1)\cdot P(E_2)$
$P(E_1|E_2) = P(E_1)$

E_1: you get even number on dice
E_2: you get H on coin
E_3: you get odd number on dice

E_1 and E_3 are mutually exclusive events

$P(E_1\cap E_2) = 0$

* **Random variable**: Quantify/Specify event outcome
  ↳ Mapping from an event to $\mathbb{R}$.

X: getting 4 or less (dice throw)

X: amount of chocolate eaten by students

X: Time spent of website

This random variable can take values $\{x_1, x_2, x_3, \dots\}$ denoted by small x

Random variable
- discrete → set of outcomes are countable
  PMF — probability Mass Function
- continuous → set of outcomes are not countable
  PDF — probability density function

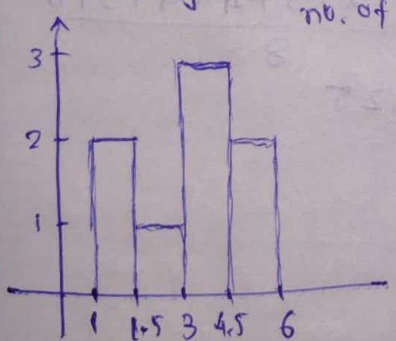
PMF ≈ PDF

X: Number of pizza sold in a day (discrete X)
↳ $X: \{1, 1, 2, 3, 4, 4, 5, 6\}$

X: diameter of Mango in inch (continuous X; it can be 2.1, 2.2223, 4.599, ... ∞)

Histogram: counting results in bin range

$$\text{bin range} = \frac{\text{max} - \text{min}}{\text{no. of bins}} = \frac{6-1}{4} = 1.5$$

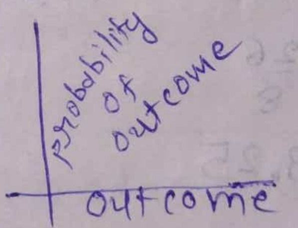


	count	probability
$1 \le x < 1.5$	⇒ 2	2/8
$1.5 \le x < 3$	⇒ 1	1/8
$3 \le x < 4.5$	⇒ 3	3/8
$4.5 \le x \le 6$	⇒ 2	2/8

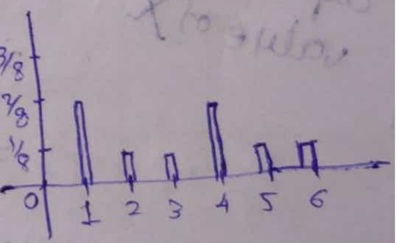
OR

x	1	2	3	4	5	6
$P(x)$	2/8	1/8	1/8	2/8	1/8	1/8

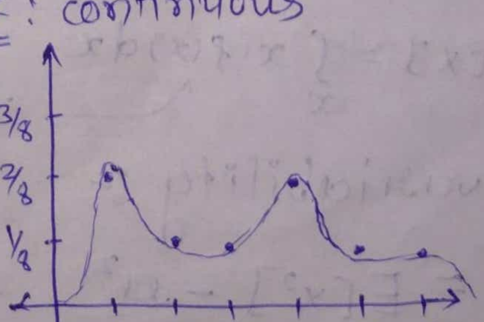
PMF: discrete



OR



PDF: continuous



value between 1 and 2 is also possible for Continuous RV X:

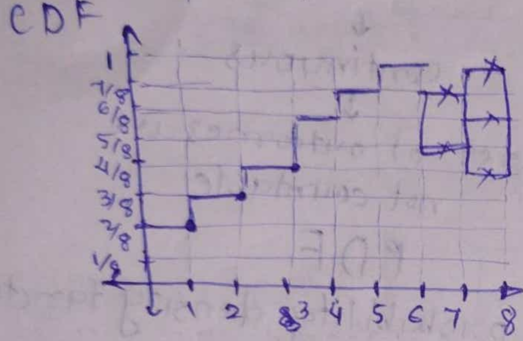
PDF or PMF = $f^{PDF}(X=a) = P(a)$ OR $P(x) = p(x)$

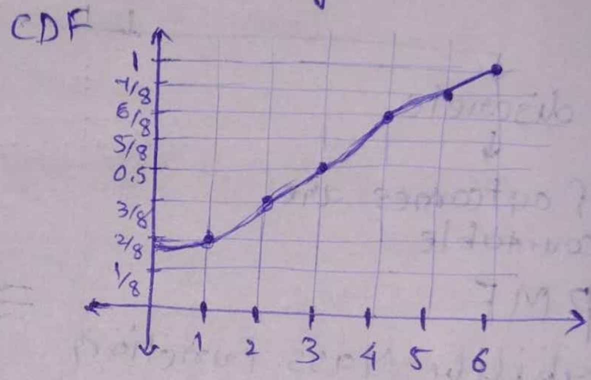
CDF cummulative distribution function $= F_X(X \le a) = F(a) = \sum_{x \le a} P(X=x) = \sum_{x \le a} p(x)$:

$x=a$	1	2	3	4	5	6
$n(x \le a)$	2	3	4	6	7	8
$p(x)$	2/8	1/8	1/8	2/8	1/8	1/8
$F(x)$	2/8	3/8	4/8	6/8	7/8	8/8

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

← cummulative probability





* **Expectation** :- what is the expected number of pizza you will sell today? (average)(μ mean)

$$E[X] = \sum_x x \cdot p(x)$$ ← PMF $P(X=x)$

Expected value of X

$$= 1\cdot\left(\tfrac{2}{8}\right) + 2\cdot\left(\tfrac{1}{8}\right) + 3\cdot\left(\tfrac{1}{8}\right) + 4\left(\tfrac{2}{8}\right) + 5\left(\tfrac{1}{8}\right) + 6\left(\tfrac{1}{8}\right)$$

$$= \frac{26}{8}$$

$$= 3.25$$

OR $\mu = \dfrac{1+1+2+3+4+4+5+6}{8}$

$= 3.25$

in case of continuous variable

$$E[X] = \int_x x \cdot f(x)\,dx.$$

PDF

* **Variance, variability** :-

$$Var(X) = E[X^2] - \mu^2$$ μ is mean of $E[X]$

OR

$$= E[X^2] - (E[X])^2$$

* **Distributions** :- distribution is nothing but to see and model how probability of an event is distributed i.e follows trend across given range.

Distribution	PDF/PMF $f(x)$	CDF $f(x) = F(X \le x)$	Expectation	Variance
Bernauli (p)	$\begin{cases} p, & x=1 \\ 1-p, & x=0 \end{cases}$	$\begin{cases} 1-p, & x \le 1 \\ 1, & x > 1 \end{cases}$	p	$pq = p(1-p)$
Binomial (n, p)	$\binom{n}{k} p^k (1-p)^k$	$\sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i}$	$n \cdot p$	$n \cdot p(1-p)$ $= npq$
UNIFORM (a, b) → continuous	$\begin{cases} \frac{1}{b-a} & x \in [a,b] \\ 0 & o.w. \end{cases}$	$\begin{cases} 0 & , x < a \\ \frac{x-a}{b-a} & , x \in [a,b] \\ 1 & , x > b \end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
→ discrete	$\begin{cases} 1/n & x \in [a,b] \\ 0 & o.w. \end{cases}$	$\begin{cases} 0 & x < a \\ i/n & x \in [a,b] \\ 1 & x > b \end{cases}$	$\frac{a+b}{2}$	$\frac{(n^2-1)}{12}$
poisson (λ)	$\frac{\lambda^k e^{-\lambda}}{k!}$	$e^{-\lambda} \sum_{i=0}^{\lfloor k \rfloor} \frac{\lambda^i}{i!}$	λ	λ
Exponential (λ) $\beta = 1/\lambda$	$\begin{cases} \lambda \cdot e^{-\lambda x} & , x \ge 0 \\ 0 & , x = 0 \end{cases}$ $Pr(X > s+t \mid X > s) = pr(X > t)$	$\begin{cases} 1 - e^{-\lambda x} & , x \ge 0 \\ 0 & , x = 0 \end{cases}$	$\beta = 1/\lambda$	$\beta^2 = 1/\lambda^2$
Normal/Gaussian (μ, σ^2)	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$	$\int_0^x f(x)\,dx$	μ	σ^2
Log-Normal (μ, σ^2)	$\frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln x-\mu}{\sigma}\right)^2}$	$\int_0^x f(x)\,dx$	$e^{\left(\mu + \frac{\sigma^2}{2}\right)}$	$(e^{\sigma^2} - 1) \cdot e^{(2\mu + \sigma^2)}$
Pareto (x_m, α) scale, shape	$\frac{\alpha \cdot x_m^\alpha}{x^{\alpha+1}}$	$1 - \left(\frac{x_m}{x}\right)^\alpha$	log normal $\xrightarrow{\log}$ normal; pareto $\xrightarrow{\text{Boxcox}}$ normal	

① For discrete RV $E(X) = x_1 P_1 + x_2 P_2 + \dots + x_n P_n$ $\quad \sum P_i = 1$, P_i is prob.

② For continuous RV $E(x) = \int_x x \cdot f(x)\,dx$ $\quad f(x)$ is prob dist fun (PDF)

$Var[X] = E[X^2] - E[X]^2$

$\sigma^2 = E[X^2] - \mu^2$