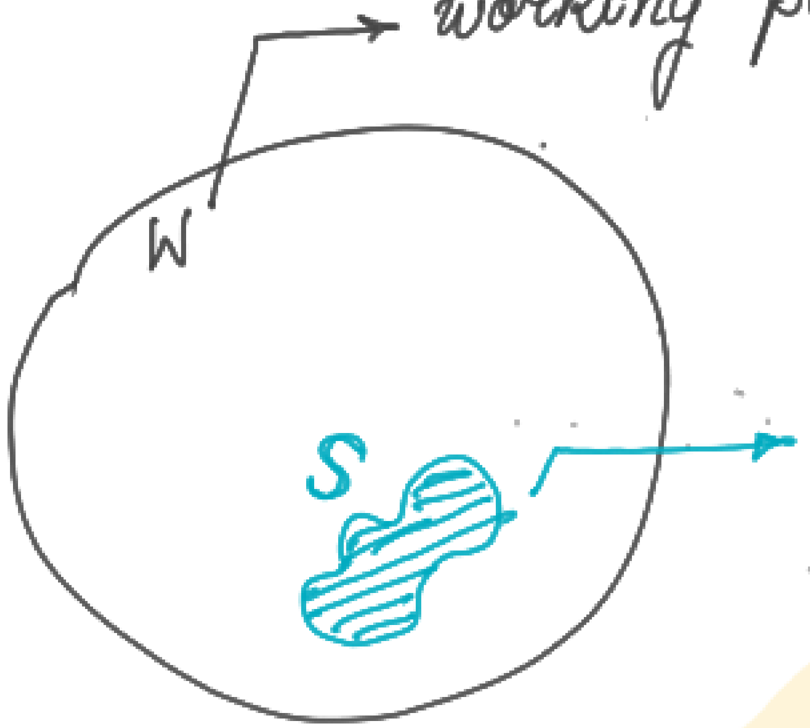
Agenda :

- Why statistics ?
- Link b/w statistics & probability
- Probability Distributions
    - Discrete
    - Continuous
- Inferential statistics
    - Central Limit Theorem

Why statistics?



Q: What is the average monthly wage of the working population of India?

very high

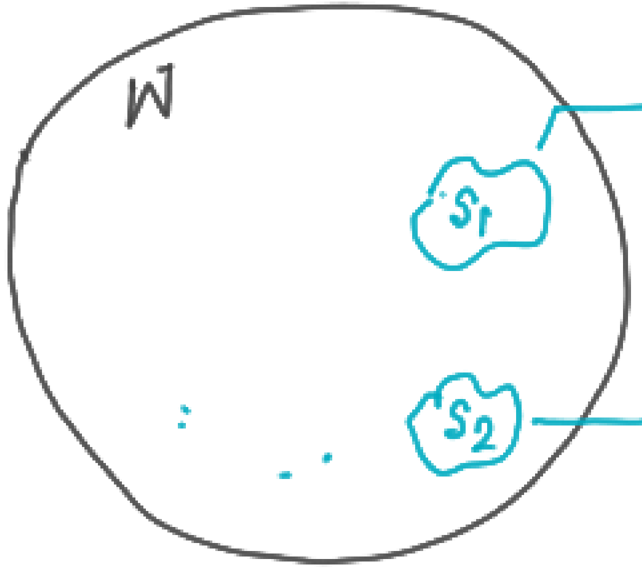
estimate

Estimation of population parameters by using information from an appropriately chosen sample.

Any sample chosen should be representative of the population

Statistics → Probability

population



sample 1 $\Rightarrow$ $\bar{\$} = 5700$

sample 2 $\Rightarrow$ $\bar{\$} = 6300$

⋮

many samples $\Rightarrow$ many values for avg. monthly wage $\Rightarrow$ How to select 'best' value?

defining 'best' $\leftarrow$

- most frequent
- middle most
- average value

can be defined in terms of probability

Random Experiment

- ↳ outcome not fixed
- ↳ set of all possible outcomes is called sample space

Example: RE: Batsman facing a bowler

SS : { 'lbw', 'no-ball', 'no-ball-4', 'no-ball-6', 'scored 4 runs', ... }

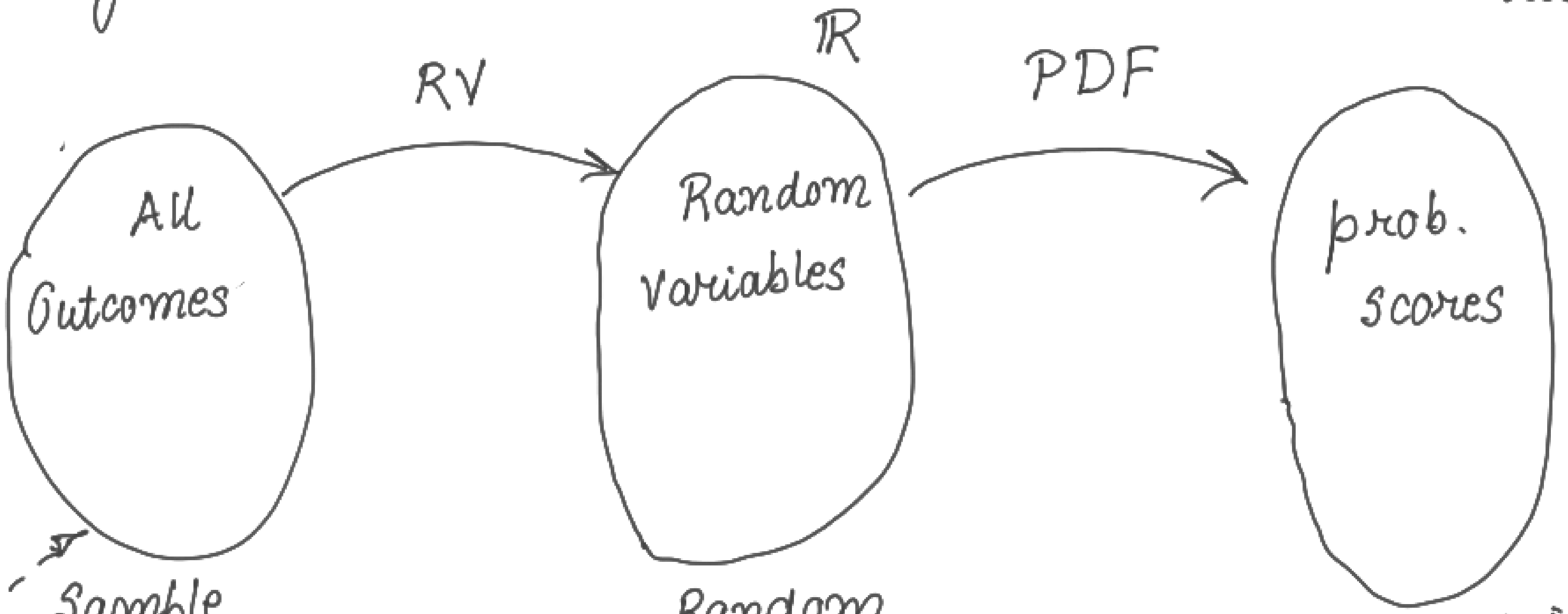
X: RV that shows # of runs scored in this ball { 0, 1, 5, 7, 4, ... }

Random variable

↳ assigns a real number to each element in the sample space

Probability Distribution

↳ assigns a number b/w [0, 1] to each value of the random variable



Sample space
- ↳ countable
- ↳ uncountable

Random Experiment

Random variables
- ↳ discrete
- ↳ continuous

probability
- ↳ discrete
- ↳ continuous

Axioms of Probability

1. $0 \leq P(X=x) \leq 1$

2. $\sum_{X} P(X=x) = 1$

Results :

1. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

2. $P(A \cap B) = P(A) \times P(B)$ if A & B are independent

3. $P(A \cap B) = \phi$ if A & B are mutually exclusive

Random Exp: Rolling an unfair dice

↳ probability of getting a number $\propto$ that number

Sample space : { ⚀ , ⚁ , ⚂ , ⚃ , ⚄ , ⚅ }

Random Variable X : # of dots on face of dice.

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

$P(X=x) \propto x \Rightarrow P(X=x) = kx \quad (k>0)$

$$\sum_{x=1}^{6} P(X=x) = 1 \Rightarrow k(1+2+3+4+5+6) = 1 \Rightarrow k = \frac{1}{21}$$

X	$P(X=x)$ unfair	$P(X=x)$ fair
1	$1/21$	$1/6$
2	$2/21$	$1/6$
3	$3/21$	$1/6$
4	$4/21$	$1/6$
5	$5/21$	$1/6$
6	$6/21$	$1/6$

$$P(X=x) = kx$$

$$k = 1/21$$

$$P(X=1) = \frac{1}{21}$$

$$\underset{\text{unfair}}{E(X)} = \sum_{x=1}^{6} x P(X=x)$$

$$= \sum x \cdot kx = k \underbrace{\sum_{1}^{6} x^2}_{91}$$

$$= \frac{91}{21} = 4.33$$

$$\underset{\text{fair.}}{E(X)} = \sum_{1}^{6} x \underbrace{P(X=x)}_{1/6}$$

$$= \frac{1}{6} \sum_{1}^{6} x = \frac{7}{2} = 3.5$$

Expected value and Variance

$$E(X) = \sum_{X} x\, P(X=x) \qquad = \int_{-\infty}^{\infty} x\, f(x)\, dx$$

↳ average value of the RV X.

$$\operatorname{var}(X) = E(X^2) - [E(X)]^2$$

↳ spread of the RV X

$$E(X^2) = \sum_{X} x^2\, P(X=x) \qquad = \int_{-\infty}^{\infty} x^2 f(x)\, dx$$

Problem (contd...)

$$E(X^2) = \sum_1^6 x^2 P(X=x)$$

unfair

$$= \sum_1^6 x^2 \cdot kx = k\sum_1^6 x^3 = k\left(\frac{6\times 7}{2}\right)^2 = \frac{1}{\cancel{21}} \times \frac{6\times\cancel{6}\times\cancel{7}\times 7}{\cancel{2}\times 2}$$

$$= 21$$

$$\underset{\text{unfair}}{var(X)} = E(X^2) - [E(X)]^2 = 21 - (4.33)^2 = 2.2$$

unfair

$$E(X^2) = \sum_1^6 x^2 \underbrace{P(X=x)}_{1/6}$$

$$= \frac{1}{6}\sum_1^6 x^2 = \frac{91}{6} = 15.16$$

$$\underset{\text{fair}}{var(X)} = 15.16 - (3.5)^2 = 15.16 - 12.25 = 2.91$$

Discrete Probability Distribution Functions

Uniform Dist. → tossing a fair dice

$X : \{1, 2, \dots k\}$

PDF : $P(X=x) = 1/k$

$E(X) = \frac{k+1}{2}$

$var(X) = \frac{k^2-1}{12}$

Bernoulli Dist. → tossing a coin / modelling binary outcome

$X : \{0, 1\}$

PDF: $p^x (1-p)^{1-x}$

$p = P(X=1)$

$E(X) = p$

$var(X) = p(1-p)$

Discrete PDF

Binomial Dist → n trials with binary outcome
x successes $(x \leqslant n)$
p is the probability of success in each trial

$X : \{0, 1, 2 \ldots n\}$

$\frac{n!}{x!(n-x)!}$

PDF : $P(X = x) = {}^nC_x \, p^x (1-p)^{n-x}$

$E(X) = np$

$var(X) = np(1-p)$

Geometric Dist → how many trials does it take to achieve first success? $p = P(X = 1)$ at each trial

$X : \{0, 1, 2, \ldots \infty)$

PDF : $P(X = x) = (1-p)^x p$

$E(X) =$

$var(X) =$

Continuous PDF

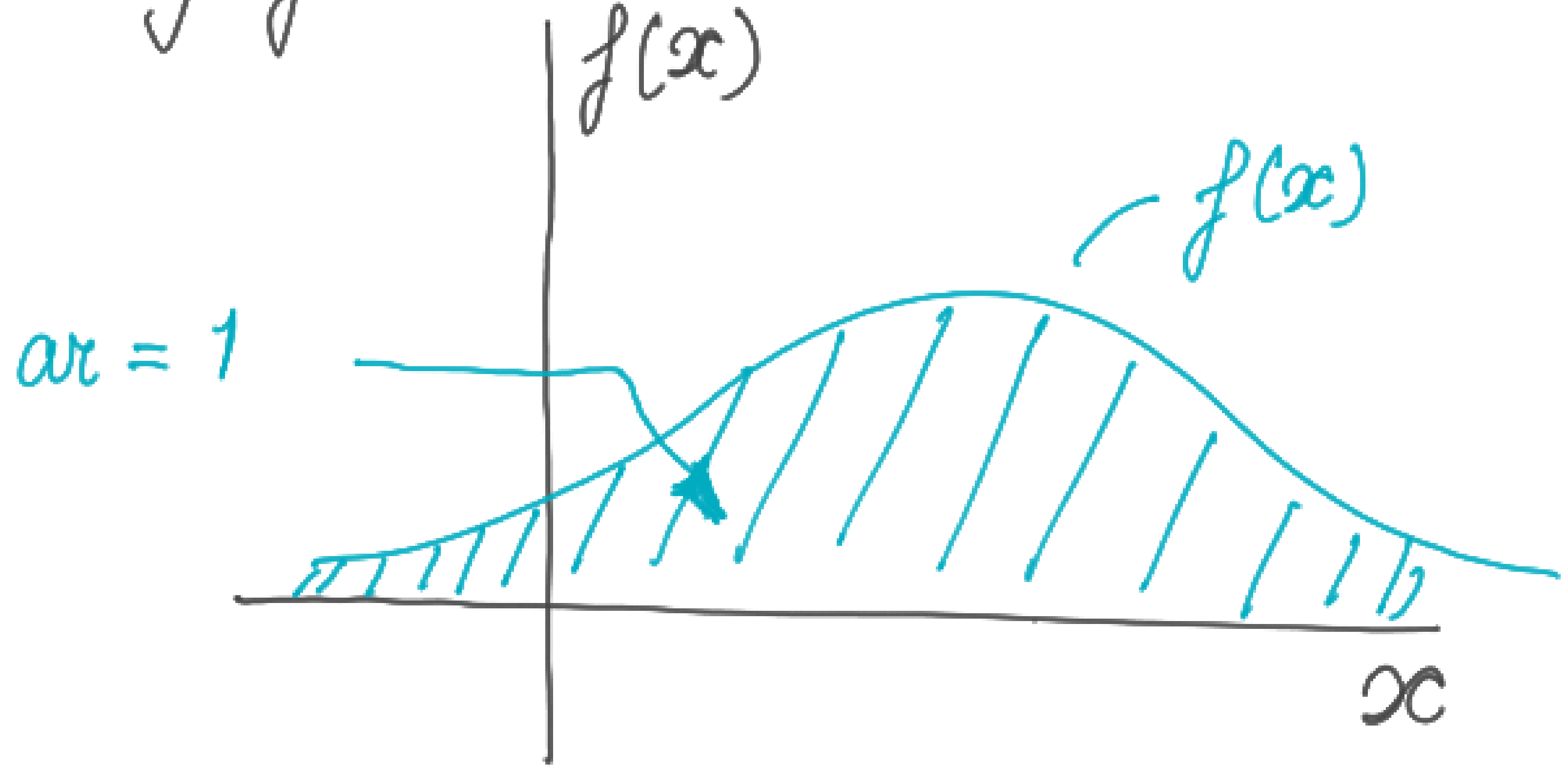
X is a continuous RV ; then $P(X = x) = 0$ always

we calculate probability at an interval

$$P(x-h < X < x+h) = \int_{x-h}^{x+h} \underbrace{f(x)} \, dx$$

probability density f^n

$$\int_{-\infty}^{\infty} f(x)\, dx = 1$$



Normal Distribution

$X : (-\infty, \infty)$

$E(X) = \mu$

$var(X) = \sigma^2$

PDF : $f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

→ used to model natural processes

Q : class of 50 students : $H \sim N(165, 25)$

$P(H < 175 \text{ cm}) = ?$ $= \int_{-\infty}^{175} \frac{1}{\sqrt{2\pi} \cdot 5} \cdot e^{-\frac{(x-165)^2}{50}} dx$

↳ hard to integrate

↳ no closed form soln; changing anything means redo calc.

Standard Normal PDF

$E(Z) = 0$

$var(Z) = 1$

$Z : (-\infty, \infty)$

PDF: $f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$

Q: $H \sim N(165, 25)$

$5Z + 165$

$P(H < 175)$

$Z = \frac{H - 165}{5}$

$P(Z < 2) = 0.5 + 0.4772 = 0.9772$

$$P(Z = 2) = \int_{-\infty}^{2} \underbrace{\frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \, dZ}_{\text{these computations are performed beforehand}}$$

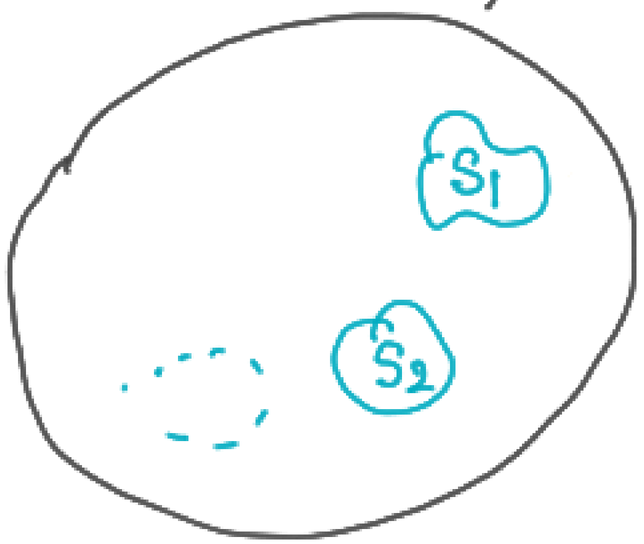
these computations are performed beforehand $\Rightarrow$ normal dist tables

Central Limit Theorem

↪ relates every distⁿ to the normal distribution

$X : RV \sim D$

population (μ, σ^2)



$$\mu = \frac{\sum_{1}^{100} \overline{X}_i}{100}$$

$$\sigma^2 = \left(\frac{n-1}{n}\right) s^2$$ → sample variance

$$s^2 = \frac{1}{100} \sum (\overline{x}_i - \mu)^2$$

$$\left.\begin{array}{l} S_1 \rightarrow \overline{X}_1 \\ S_2 \rightarrow \overline{X}_2 \\ S_3 \rightarrow \overline{X}_3 \\ \vdots \qquad \vdots \\ S_{100} \rightarrow \overline{X}_{100} \end{array}\right\} \quad \overline{X}_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

n → sample size