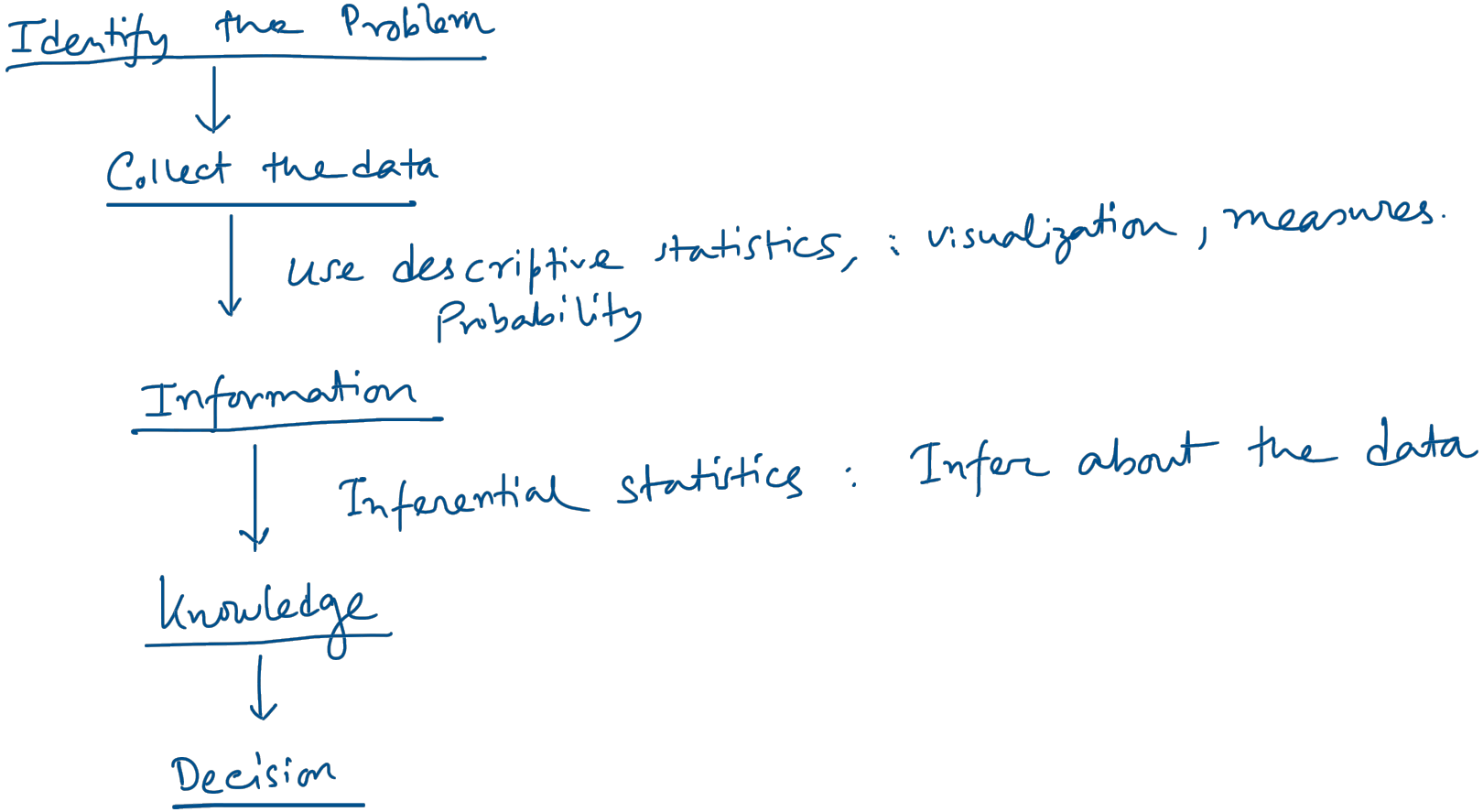
Inferential Statistics

Statistical Decision Making Process



Statistics

Descriptive Statistics

1. Summarization, presentation visualization of the data
2. mean, median, mode
   range, variance, IQR
3. Gain the insights from the data.

Inferential Statistics

1. Involves making predictions or inferences about a larger population based on the observed statistics in the sample data.
2. Involves techniques:
   Hypothesis testing
   Confidence interval

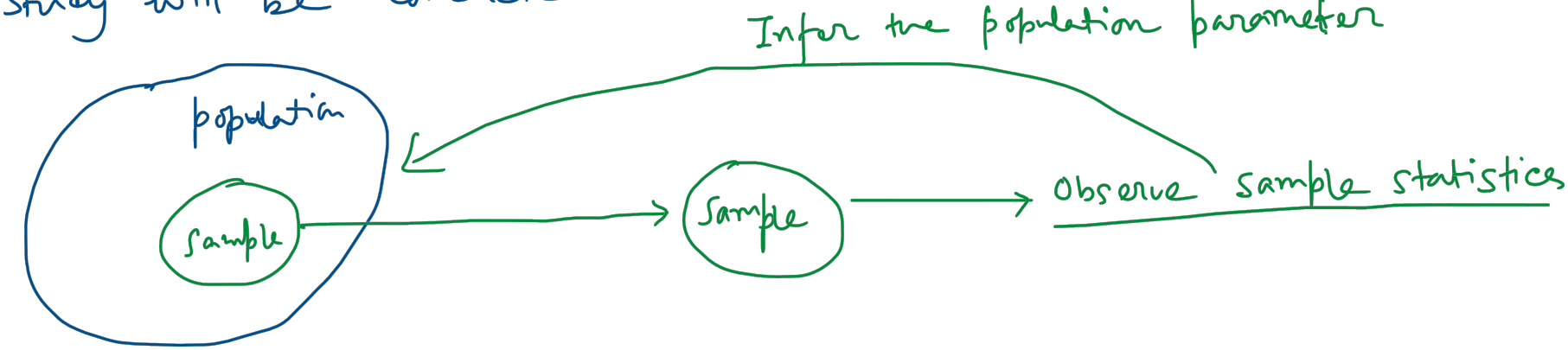
Population

1. The entire group that we want to study

examples: (i) Average height of babies between the age 1 - 5 years in a particular region : [all the babies in that region age between 1 to 5]

Sample

2. A smaller, representative subset of the population on which the study will be conducted.



Population

Parameters

- population mean / proportion
- Population standard deviation

Sample

Statistics

- Sample mean / proportion
- Sample std. deviation

Sample statistics provides an estimate of the population parameters.

There are different sampling techniques

1. Simple Random Sampling (SRS)
   - with replacement ✓
   - without replacement ✓

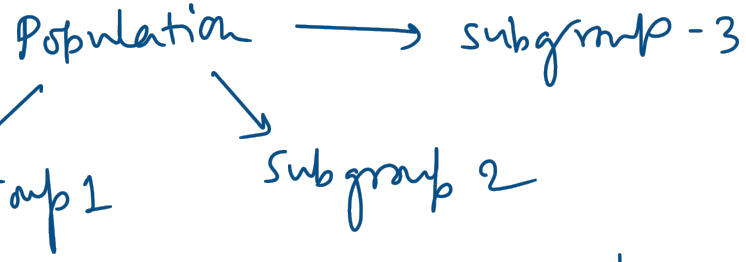
N = 1000

rand(N) → 1 to 1000

100 , 100 .....

2. Stratified random sampling

The population is divided into subgroups called strata



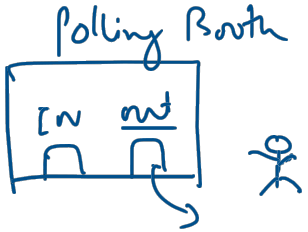
Now from each strata simple random sample are taken in proportion to the population distibution.

Adults in Town

- young adults : 18 Yo – 25 Yo [25 %] → 25 → 28
- mid age adults : 25 Yo – 35 Yo [15 %] → 15 → 12
- 35 Yo – 50 Yo [15 %] → 15 → 16
- 50 Yo – 65 Yo [20 %] → 20 → 25
- 65 + Yo [25 %] → 25 → 20

100

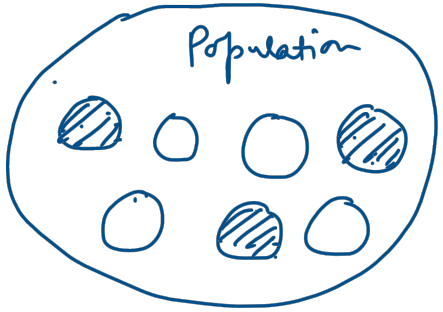
3. **Systematic Sampling**: Individuals are taken at regular interval from a randomly chosen starting point.



10, 20, 30, 40, ...

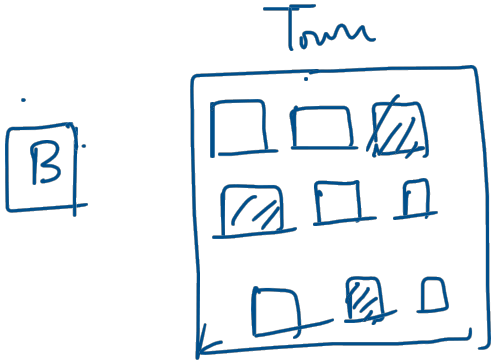
Every 10th person coming out from the poling booth.

4. **Cluster Sampling**:



N - clusters.

5 cluster out of the N cluster



5. Convenience Sampling

6. Snowball Sampling

7. Purposive Sampling

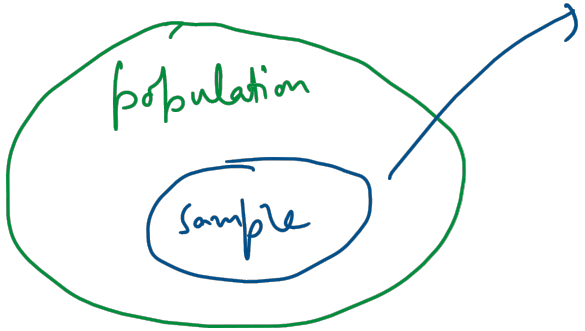
Population → Sample

Population:
- population mean (μ)
- population std. deviation (σ)
- / population proportion (p)

Sample:
- Sample mean ($\bar{x}$)
- Sample std. deviation (s)
- / Sample proportion ($\hat{p}$)

$\bar{x}$ is an estimate of μ., $\hat{p}$ is an estimate of p and s is an estimate of σ.

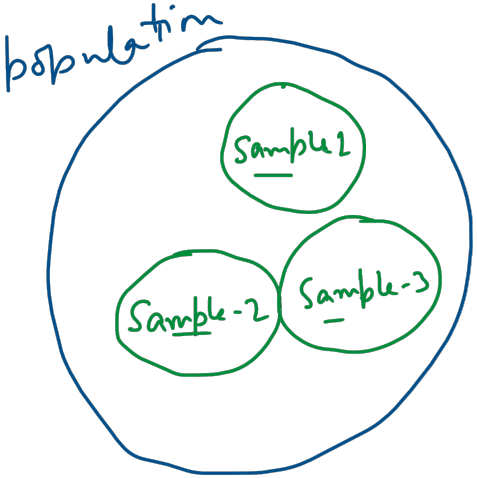
Estimator and Estimates:



Estimator provides an estimate of the certain population parameter solely based on the sample collected.

I want to estimate population mean (μ)

I have sample mean ($\bar{x}$) : an estimator of population mean.

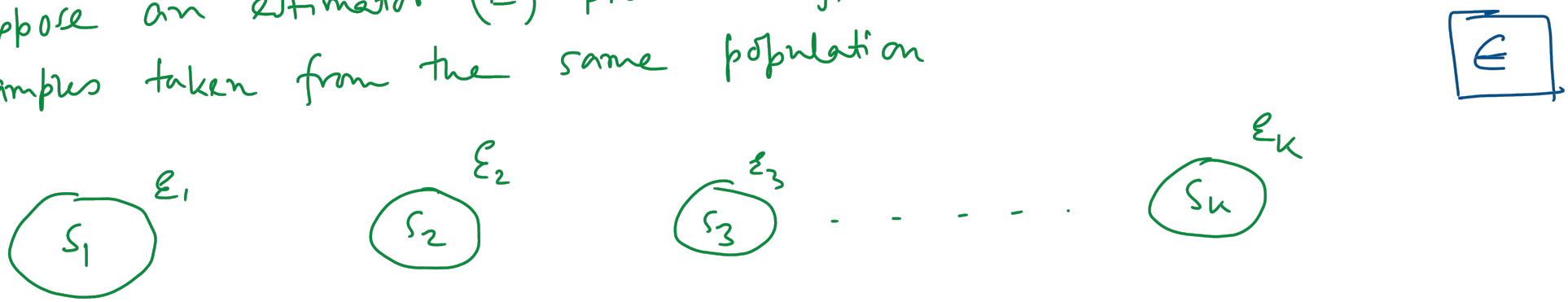


Samples	$\bar{x}$
S_1	$\bar{x}^{(1)}$
S_2	$\bar{x}^{(2)}$
S_3	$\bar{x}^{(3)}$
S_n	$\bar{x}^{(n)}$

What properties should an estimator have so that it is called a good estimator?

→ No Estimator is perfect

→ Suppose an estimator (ϵ) produces different estimates of different samples taken from the same population



$$\mathbb{E}[\underline{\epsilon}] = \frac{1}{k}(\epsilon_1 + \epsilon_2 + \cdots + \epsilon_k) \approx \underline{\epsilon}$$

$\mathbb{E}[\epsilon] = \epsilon$: Then ϵ is an unbiased estimator

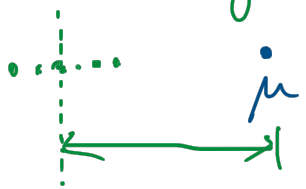
A good estimator on an average gives a very good estimate of the population parameter.

Each Estimator is associated with Bias and Variance

E_1 : low bias / high variance

$\mathbb{E}[\bar{x}]$ $\quad$ $\text{var}(\bar{x})$

E_2 high bias / low variance



E_3

low bias / low variance

E_4 high bias / high variance

$\mathbb{E}[\bar{x}] - \mu$: bias

$\text{Var}(\bar{x})$: variance

Sample Mean :

Suppose the sample is drawn from the population and sample size is 'n'.

$x_1, x_2, x_3, \ldots, x_n$ then $\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$

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

x_i are independent and identically distributed random variables with mean = population mean (μ) and std. deviation = σ (pop std. deviation)

$x_i \sim (\mu, \sigma^2)$
i.i.d.

$$E[\bar{x}] = E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[x_i]$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{n\mu}{n} = \mu$$

$E[x_i] = \mu$

$Var(x_i) = \sigma^2$

$E[\bar{x}] = \mu$ → Unbiased Estimator of population mean.

$$Var(\bar{x}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2} Var\left(\sum_{i=1}^{n} x_i\right)$$

$$\therefore Var(\bar{x}) = \frac{1}{n^2} \cdot \sum_{i=1}^{n} Var(x_i)$$

$$= \frac{1}{n^2} \sum_{i=1}^{n} \sigma^2$$

$$= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

$$Var(\bar{x}) = \frac{\sigma^2}{n} \quad \text{and} \quad std(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

sample size ↑ → variance($\bar{x}$) ↓

$std(\bar{x})$ → **Standard Error**

$x \rightarrow \sigma$

$Y = aX$

$Var(Y) = a^2 Var(X)$

$Z = X + Y$

$Var(Z) = Var(X) + Var(Y)$

n : sample size

sample size ↑ → standard error ↓

Central Limit Theorem :-

$X_1, X_2, \ldots, X_n$ — n independent and identically distributed random variable with mean μ and std. deviation σ

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \bar{x} \rightarrow \text{sample mean}$$

as $n \rightarrow \infty$ the $\bar{X}_n \rightarrow N\left(\mu, \frac{\sigma^2}{n}\right)$

As the sample size increases the sample mean approaches towards a normal distribution with mean $= \mu$ and Std. deviation $= \frac{\sigma}{\sqrt{n}}$

This is called **sampling distribution**

The sampling distribution $\sim N\left(\mu, \frac{\sigma^2}{n}\right)$ no matter what the original distribution is

Sample Standard deviation:

$x_1, x_2, \ldots, x_n$: sample values. , $\bar{x}$ = sample mean $= \frac{1}{n}\sum_{i=1}^{n} x_i$

Sample Std. deviation (s) $= \sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

Why $(n-1)$? and not n?

Population std. deviation (N values)

μ: pop. mean

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

Suppose we define sample std. deviation

$\hat{s}_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$, then $\hat{s}_n$ is a biased estimator of σ

$\hat{s}_n$ always under estimates σ.

Dof : Degrees of freedom

x_1, x_2, x_3, x_4, x_5

$$\bar{x} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5}$$

if I fix $\bar{x}$ and change $x_1, x_2, \ldots, x_5$ then how many of these 5 values can we change?

$x_1 = 2$

$x_2 = 3$

$x_3 = 1$

$x_4 = 5$

$x_5 = 3$

$\bar{x} = \frac{14}{5} = 2.8$

$x_1 = 3$

$x_2 = 1$

$x_3 = 5$

$x_4 = 2$

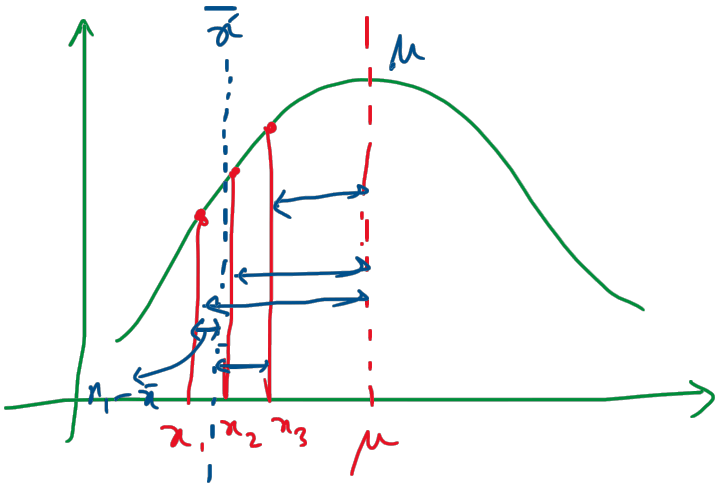
$\Rightarrow x_5 = 3$

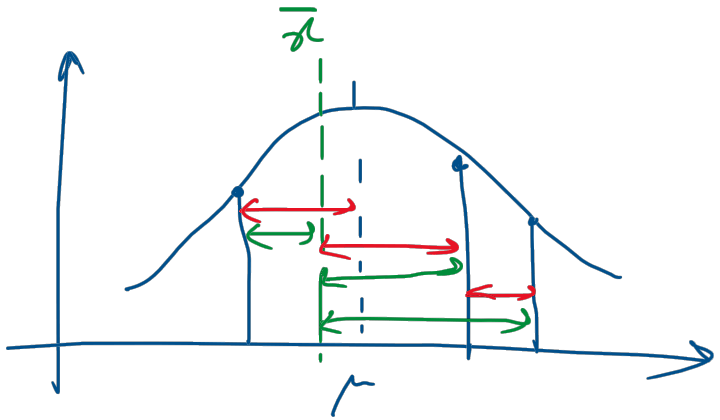
(4)

(n) $\bar{x}$

n − 1 $\bar{x}$

n − dof = ...





$$E\left[\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{n}\right] > E\left[\sum_{i=1}^{n}\frac{(x_i-\bar{x})^2}{n}\right]$$

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