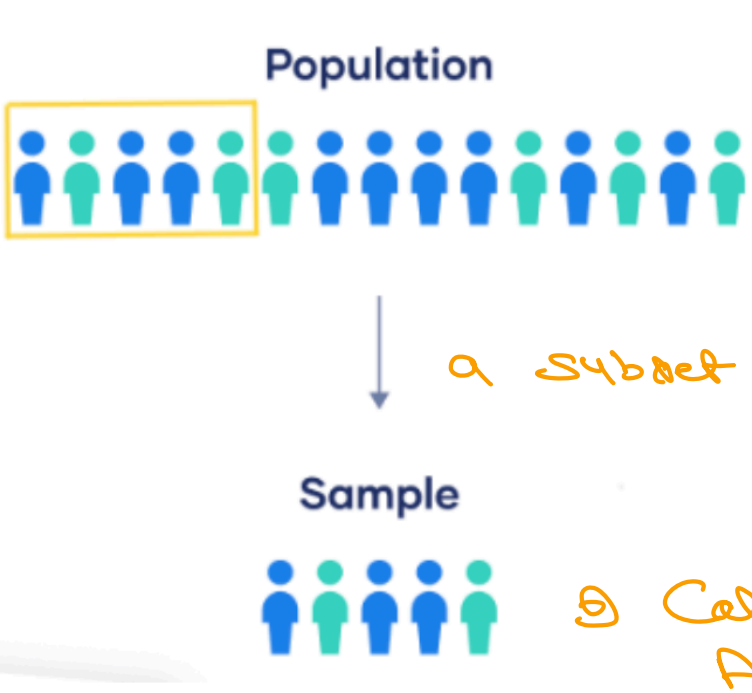
* Population vs Sample
* Sample Statistics
  - Mean
  - Variance
  - SD
* Point Estimates
* Sampling Distribution
* Standard Error
* Uniform Distribution
  - PMF
  - PDF
  - CDF

Population vs Sample

Goal: Find average income of ALL M2 Professional in BLR

* Ideally: we would like to collect all the data possible
  * Not feasible
  * Very costly



Most practical Scenario

a subset of → population

Calculate statistics Avg Salary with some error margin

Sample Statistics

Population Stats (N)	Sample Stats (n)
$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
$\sigma^2 = \frac{\sum_{i=1}^{N}(\mu - x_i)^2}{N}$	$s^2 = \frac{\sum_{i=1}^{n}(\bar{x} - x_i)^2}{n-1}$ s^2: Sample Variance $n-1$: Bessel's Correction (Why ?)
$\sigma = \sqrt{\sigma^2}$	$s = \sqrt{\frac{\sum_{i=1}^{n}(\bar{x} - x_i)^2}{n-1}} = \sqrt{s^2}$

Why do we need Sample Statistics

Analyze population

↓

Sample Statistics → estimate → Population Parameters

Quiz Ended!

There are 45 students in a class.
5 students were randomly selected from this class and their heights (in cm) were recorded as follows:
[131, 150, 140, 142, 152] Calculate Sample mean and sample variance

143

$$\frac{(131-143)^2 + (150-143)^2 - \cdots}{5-1}$$

$$(\bar{x} - x_1)^2$$

$$(x_i - \bar{x})^2$$

Estimate

- Point Estimate → 20 Lakhs
- Range Estimate → 18 – 22 Lakhs, 95%

Sampling Technique

① Probabilistic Sampling

Every single member of population will have a probability assigned for it to be part of Sample

② Non-probabilistic Sampling X

- Generally based on convenience and Hence cheaper.
- Convenience Sampling
- Snowball Sampling

There are four main types of probability samples.

1. **Simple random sampling** (Today)
2. **Systematic sampling**
3. **Stratified sampling**
4. **Cluster sampling**

ML Module → Supervised, Unsupervised, RL

P → 1/10,000

10,000 ML B Population Size

Sampling : Margin of Error

Case-1 : 0% Margin of Error ⇒ 10000

Case-2 : 5% Margin of Error ⇒ $n_1 < 10000$

Case-3 : 10% Margin of Error ⇒ $n_2 < n_1$

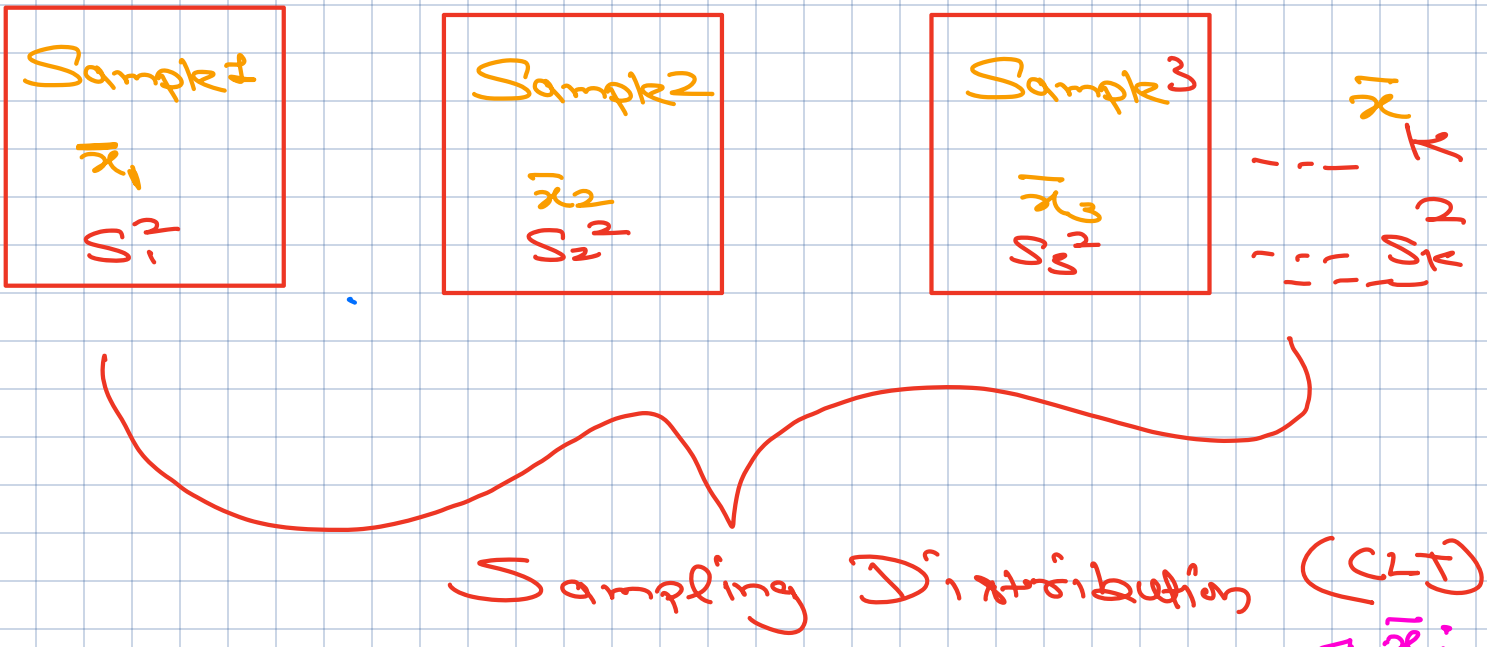
* In Random Sampling all points will have equal probability of being picked

10000 people

Steps for Random Sampling

1. Define Population and sample-Size
2. Pick Sampling technique and perform Sampling
3. Collect Sample Data → ML Model
   → Statistical Analysis

Multiple Samples



Sampling Distribution (CLT) $\Rightarrow \bar{x}_i$

(Sample by Sample)

* Estimate Population Parameter with Higher reliability (CLT)

(1) $S_1 \rightarrow$ 22.8 $\bar{x}_1$

(1) $S_2 \rightarrow$ 23.8 $\bar{x}_2$

(1) $S_3 \rightarrow$ 21.2 $\bar{x}_3$

CLT

$[\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots \bar{x}_k]$

Sampling Distribution

Standard Error

↓

To quantify variability across Sample

↓

How far population mean is from Sampling Distribution mean

① Population SD is known

$$SE = \frac{\sigma}{\sqrt{n}}$$

$\sigma \rightarrow$ Population SD
$n \rightarrow$ Sample size

② Population SD Not known

$$SE = \frac{s}{\sqrt{n}}$$

s - Sample SD
$n \rightarrow$ Sample size

* If SE is large : Pop-mean is far away from Sample-mean

* If SE is Low : Pop-mean is Close from Sample-mean

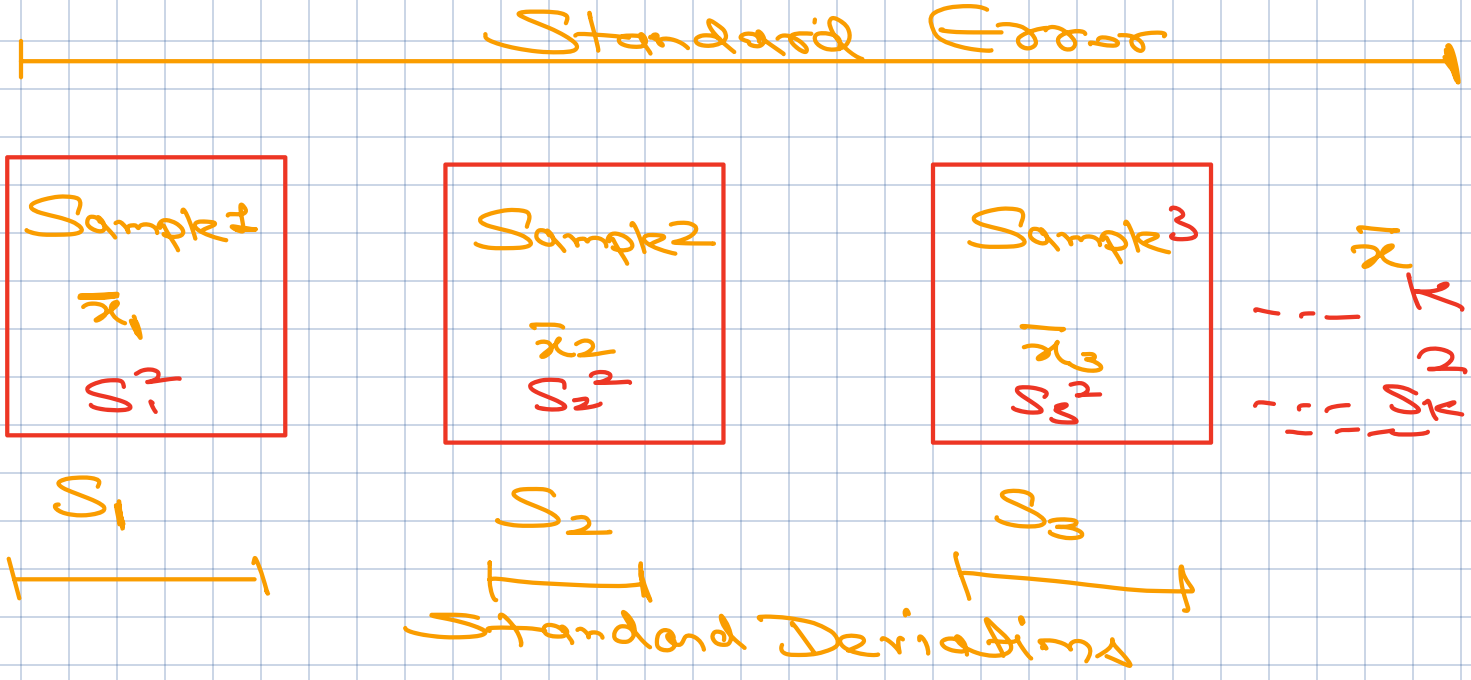
Key Takeaway

Since SE $\propto \frac{1}{\text{Sample size}}$

Higher Sample Size, leads to the Standard Error

★ Law of Large Numbers

- Higher the Sample size, the close Sample mean gets to Population Mean.



Collection of Sample mean → (Sampling Distribution)

SE →

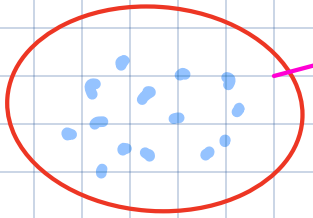
A sample of 30 latest returns on XYZ stock reveals a mean return of 4 with a sample standard deviation of 0.13. Estimate the SE of the sample mean.

$n = 30$

$\bar{x} = 4$

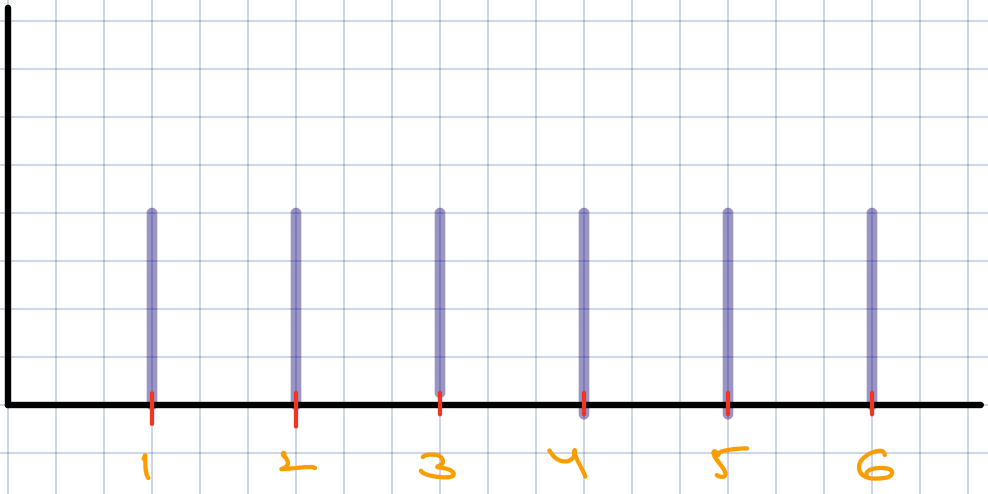
$S = 0.13$

$SE \rightarrow \frac{0.13}{\sqrt{30}}$

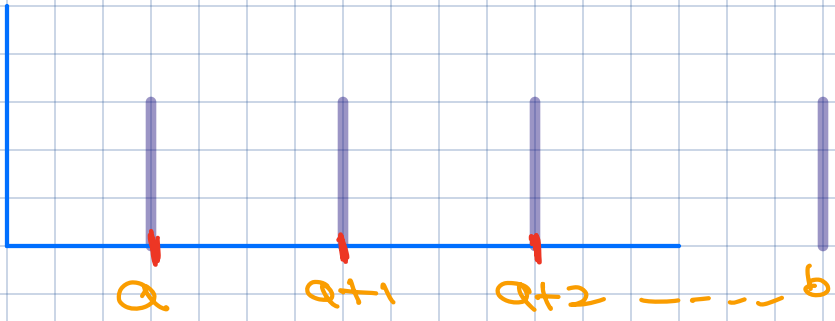


Uniform Distribution

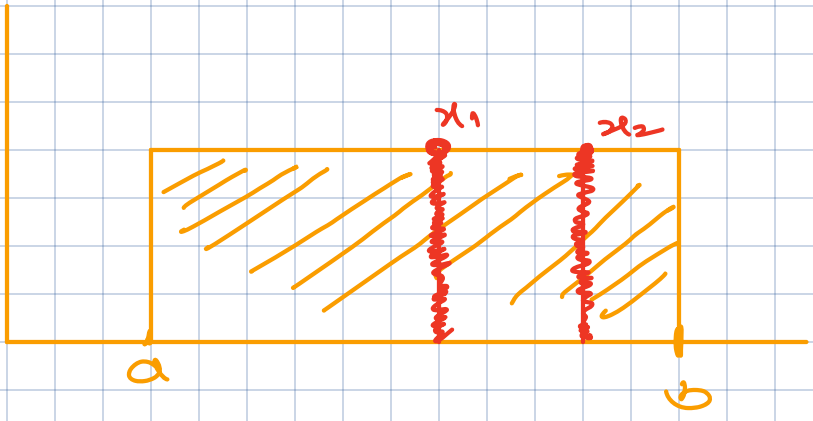
A distribution where all points are equally likely to occur



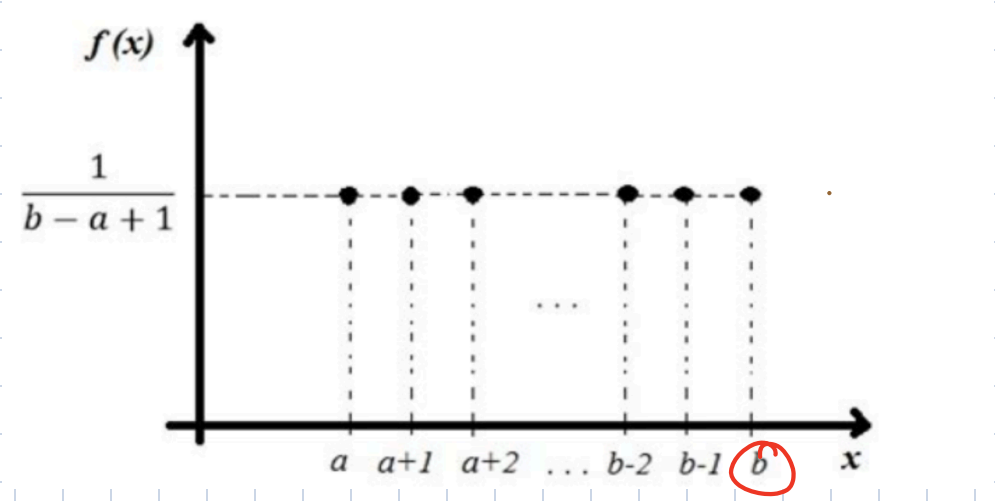
① Discrete Uniform Distribution



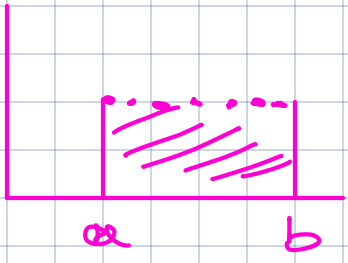
② Continous Uniform Distribution



PMF = $\frac{1}{(b-a+1)}$



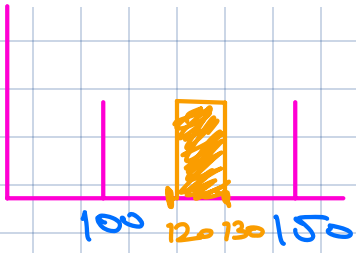
* PDF = $\frac{1}{b-a}$



Suppose the weight of dolphins is uniformly distributed between 100 pounds and 150 pounds. If we randomly select a dolphin at random, then determine the probability that the chosen dolphin will weigh between 120 and 130 pounds.

130 − 120 = 10

4 options



$\left(\frac{1}{b-a}\right)$ $\frac{1}{50} \times 10$

Check if random.uniform actually generates data from UD

```
import numpy as np
import matplotlib.pyplot as plt

# Generate random data following a uniform distribution
data = np.random.uniform(1, 7, 10000)  # 10,000 samples from a uniform distribution between 1 and 6

# Create a histogram
plt.hist(data, bins=6, edgecolor='black', align='mid', rwidth=0.8, density=True, color='skyblue')

# Set labels and title
plt.xlabel('Outcome')
plt.ylabel('Probability')
plt.title('Histogram of a Uniform Distribution (Die Roll Simulation)')

# Show the plot
plt.grid(True)
plt.show()
```

PDF $= \frac{1}{b-a}$ $\frac{1}{150-100} = \frac{1}{50}$