Random Sampling and Sampling Distribution

Objective :

To explore random sampling from a population and understand the concept of sampling distribution using Python in Jupyter Notebook.

Steps :

1. Generate a Population :

    ⟹ Create a population of data with a specified distribution (Eg - normal distribution)

2. Random Sampling :

    ⟹ Perform random sampling from population to create multiple samples of different sizes.

    ⟹ Compute sample statistics (mean, standard deviation, etc) for each sample.

3. Sampling Distribution :

    ⟹ Plot histograms or density plots of sample statistics (Eg - sample means)

    ⟹ Compare the sampling distribution of sample statistic (mean) with population distribution.
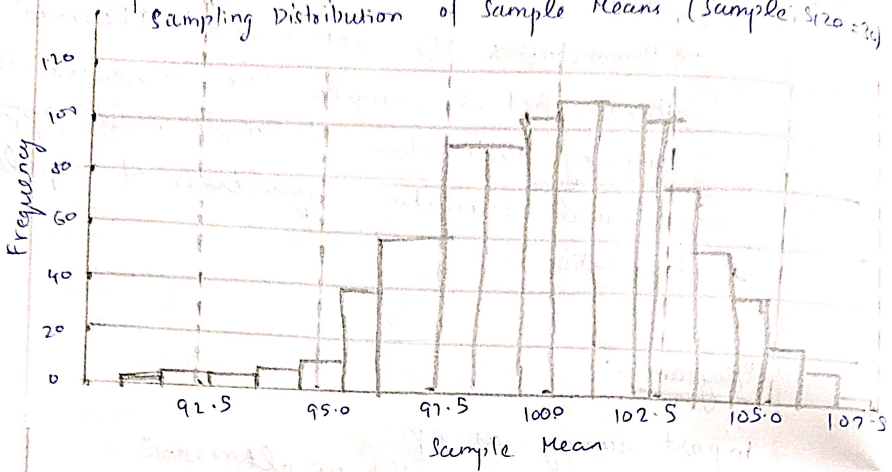
4. Central Limit Theorem (Optional):

=> Demonstrates the central limit theorem by showing that as sample size increases the sampling distribution of sample mean approaches a normal distribution regardless of population distribution.
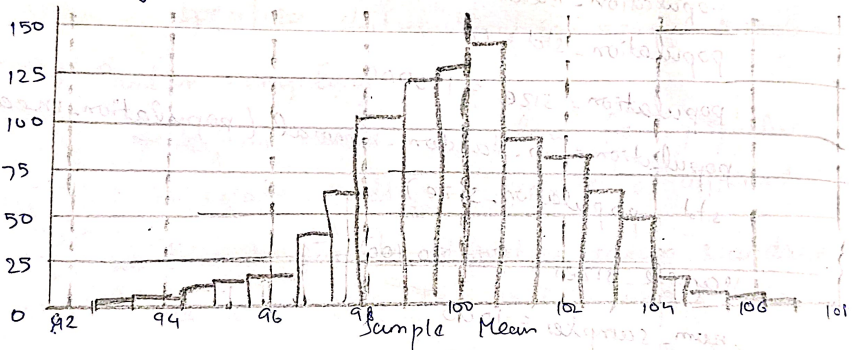
Program:

```
import numpy as np
import matplotlib.pyplot as plt
population_mean = 50
population_std = 10
population_size = 100000
population = np.random.normal(population_mean, population_std, population_size)

sample_sizes = [30,50,100]
num_samples = 1000
sample_means = {}
for size in sample_sizes:
    sample_means[size] = []

    for _ in range(num_samples):
        sample = np.random.choice(population, size=size, replace=False)
        sample_means[size].append(np.mean(sample))

plt.figure(figsize=(12,8))
```
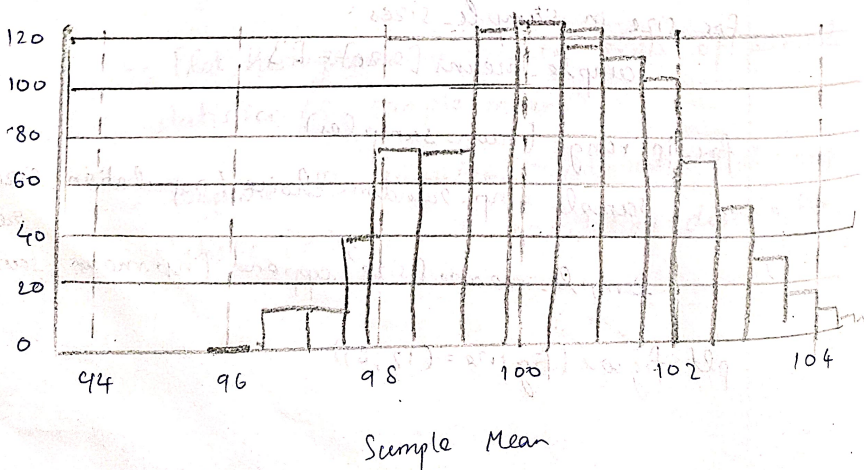
Output:

Sampling Distribution of Sample Means (Sample Size = 30)

Frequency axis: 0, 20, 40, 60, 80, 100, 120

Sample Mean axis: 92.5, 95.0, 97.5, 100.0, 102.5, 105.0, 107.5

Sampling Distribution of Sample Means (Sample Size = 50)

Y axis: 0, 25, 50, 75, 100, 125, 150

Sample Mean axis: 92, 94, 96, 98, 100, 102, 104, 106, 108

Sampling Distribution of Sample Mean (Sample Size = 100)

Y axis: 0, 20, 40, 60, 80, 100, 120

Sample Mean axis: 94, 96, 98, 100, 102, 104

```python
for i, size in enumerate(sample_sizes):
    plt.subplot(len(sample_sizes), 1, i+1)
    plt.hist(sample_means[size], bins=30, alpha=0.7,
        label=f'Sample Size {size}')
    plt.axvline(np.mean(population), color='red',
                linestyle='dashed', linewidth=1.5,
                label='Population Mean')
    plt.title(f'Sampling Distribution (sample size{size})]
    plt.xlabel('Sample Mean')
    plt.ylabel('Frequency')
    plt.legend()
    plt.tight_layout()
    plt.show()
```

Result :

The sampling distributions of sample means approached a normal shape as the sample size increased, verifying the Central Limit Theorem (CLT).