Sampling Distribution

A Sampling Distribution is a probability distribution of a statistic obtained through a large number of samples drawn from a specific population. For example, consider a normal population with mean μ and variance σ. Assume we repeatedly take samples of a given size from this population and calculate the arithmetic mean for each sample. This statistic is then called the sample mean. Each sample has its own average value, and the distribution of these averages is called the "sampling distribution of the sample mean.

A Sampling Distribution behaves much like a normal curve and has some interesting properties:

- The shape of the Sampling Distribution does not reveal anything about the shape of the population.
- Sampling Distribution helps to estimate the population statistic, using the Central Limit Theorem

Let's illustrate Sampling Distribution in Python.

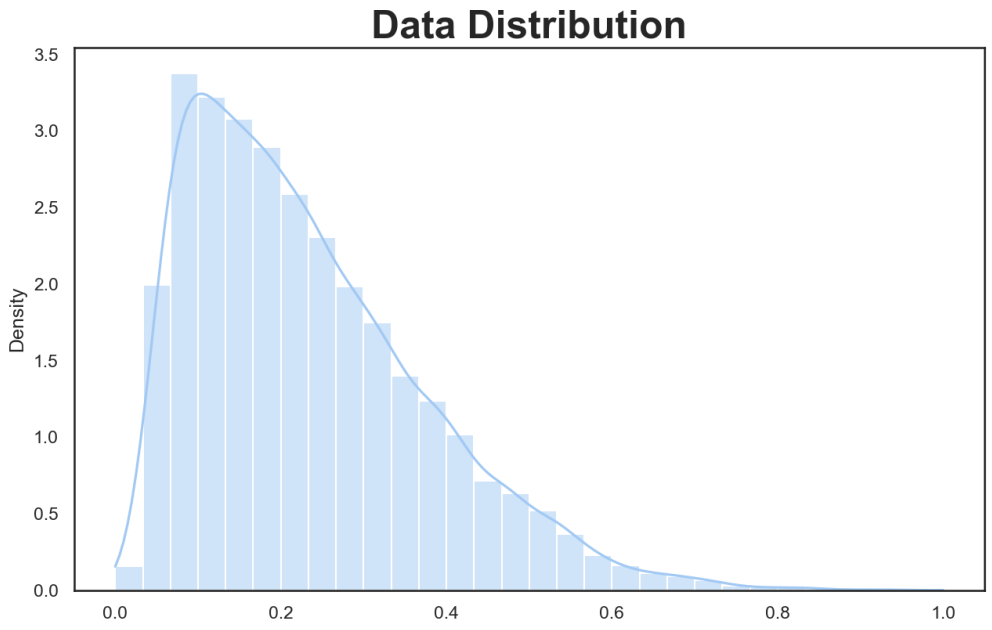
Note: Here we are using the seaborn and scipy library to visualize the data, this has not yet been introduced in the course. It will be taught later.

Let's first generate random skewed data that will result in a non-normal (non-Gaussian) data distribution. The reason behind generating non-normal data is to better illustrate the relation between data distribution and the sampling distribution.

So, let's import the Python plotting packages and generate right-skewed data.

```
# Plotting packages and initial setup
import matplotlib.pyplot as plt
import matplotlib as mpl
# Generate Right-Skewed data set
import seaborn as sns
from scipy.stats import skewnorm
from sklearn.preprocessing import MinMaxScaler

sns.set_theme(palette="pastel")
sns.set_style("white")

mpl.rcParams["figure.dpi"] = 150
```

```
num_data_points = 10000
max_value = 100
skewness = 15   # Positive values are right-skewed

skewed_random_data = skewnorm.rvs(a = skewness,loc=max_value,
size=num_data_points, random_state=1)
skewed_data_scaled =
MinMaxScaler().fit_transform(skewed_random_data.reshape(-1, 1))

# Plot the data (population) distribution
fig, ax = plt.subplots(figsize=(10, 6))
ax.set_title("Data Distribution", fontsize=24, fontweight="bold")

sns.histplot(skewed_data_scaled, bins=30, stat="density",
kde=True, legend=False, ax=ax)
plt.show()
```



Now we are going to use sampling distribution on this data:

Sampling distribution: The frequency distribution of a sample statistic (aka metric) over many samples drawn from the dataset[1]. Or to put it simply, the distribution of sample statistics is called the sampling distribution.

The algorithm to obtain the sampling distribution is as follows:

1. Draw a sample from the dataset.
2. Compute a statistic/metric of the drawn sample in Step 1 and save it.

3. Repeat Steps 1 and 2 many times.
4. Plot the distribution (histogram) of the computed statistic.

Let's plot in python:

```
import numpy as np
import random

sample_size = 50
sample_mean = []

random.seed(1)  # Setting the seed for reproducibility of the result
for _ in range(2000):
    sample = random.sample(skewed_data_scaled.tolist(), k=50)
    sample_mean.append(np.mean(sample))

print(f"Mean: {np.mean(sample_mean)} \n")

# Plot the sampling distribution
fig, ax = plt.subplots(figsize=(10, 6))
ax.set_title("Sampling Distribution", fontsize=24, fontweight="bold")

sns.histplot(sample_mean, bins=30, stat="density", kde=True, legend=False)
plt.show()
```